Author(s): Neittaanmäki, Reeta; Lamprianou, Iasonas

Title: Communal factors in rater severity and consistency over time in high-stakes oral assessment

Year: 2024

Version: Published version

Please cite the original version:

Neittaanmäki, R., & Lamprianou, I. (2024). Communal factors in rater severity and consistency over time in high-stakes oral assessment. Language Testing, OnlineFirst. https://doi.org/10.1177/02655322241239363

Original Manuscript

LANGUAGE TESTING

Language Testing
1–22

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322241239363
journals.sagepub.com/home/ltj
Sage

Communal factors in rater severity and consistency over time in high-stakes oral assessment

Reeta Neittaanmäki
University of Jyväskylä, Finland

Iasonas Lamprianou
University of Cyprus, Cyprus

Abstract

This article focuses on rater severity and consistency and their relation to major changes in the rating system in a high-stakes testing context. The study is based on longitudinal data collected from 2009 to 2019 from the second language (L2) Finnishspeaking subtest in the National Certificates of Language Proficiency in Finland. We investigated whether rater severity and consistency changed over that period and whether the changes could be explained by major changes in the rating system, such as the change of lead examiner, the modus of rating and training (on-site or remote), and the composition of the rater group. The data consisted of 45 rating sessions with 104 raters and 59,899 examinees and were analysed using the Many-Facets Rasch model and generalized linear mixed models. The analyses indicated that raters as a group became somewhat more lenient over time. In addition, the results showed that the rater community and its practices, the lead examiners, and the modus of rating and training can influence the rating behaviour. Finally, we elaborate on implications for both research and practice.

Keywords

Generalized linear mixed model, lead examiner, Many-Facets Rasch model, on-site rating, rater consistency, rater severity, remote rating, speaking assessment

Introduction

It has long been maintained that high-stakes public examinations are established and used under the assumption of trust and support by societies (Simpson & Baird, 2013;

Corresponding author:
Reeta Neittaanmäki, Centre for Applied Language Studies, University of Jyväskylä, PO Box 35, FI-40014, Finland.
Email: reeta.neittaanmaki@jyu.fi

William, 1996). For examinations, however, there is a need to produce evidence regarding their validity and reliability coupled with adequate justification of score interpretations. In high-stakes language assessments, where humans rate speech performances, it is imperative to monitor the quality of ratings to ensure that all examinees are judged fairly and that their test scores are comparable.

Human raters, however, bring subjectivity and undesirable variance to the rating process because they often make different or even conflicting judgements about examinees' language ability. In the pertinent literature, this phenomenon has been termed "rater effects" and has been attributed to a large variety of factors, such as rating experience (Davis, 2016; Lim, 2011), interpretation and the application of rating criteria (Barkaoui, 2010; Weigle, 2002), teaching experience (Bonk & Ockey, 2003), and the cohesion of the group of raters (Lamprianou et al., 2021). Solving the problem would be straightforward if assessment researchers/practitioners could identify and remove the repeatedly "misbehaving" raters from the rating process. This is not always feasible in practice, however, especially in contexts such as the one described in this study, where rating occurs within a short period of time.

Indeed, many researchers have found that humans' rating behaviour is not necessarily stable over time. For example, Congdon and McQueen (2000) found significant fluctuations of rating behaviour on different days. Hoskens and Wilson (2001) showed that raters' severity changed over different rating periods, longer than a day. Lamprianou et al. (2021) found that rating severity and consistency could fluctuate in successive years. But why would this be surprising? If, for example, severity and consistency are affected by rating experience (as the literature has shown), rating behaviour should change over time, as raters participate in successive rating sessions and accumulate experience. Although this argument may seem sound, other researchers (e.g., Leckie & Baird, 2011; Lim, 2011; Lunz & Stahl, 1990) have found that rating characteristics may be stable over time.

These conflicting findings have led to what Lamprianou (2018, p. 431) has described as the "confusion" and "agony" of policymakers, practitioners, and researchers. The phenomenon is complex and difficult to address if appropriate data are not available. For example, Lamprianou et al. (2021) suggested that "studies using operational data may enjoy a high degree of external validity" and encouraged researchers "to invest in longitudinal designs which span across substantial periods of time (e.g., many months or years)" (p. 292). Although real-life, operational data are messy and difficult to handle—see for example the travails of Lim (2011) and Myford and Wolfe (2009) to establish "connectivity" in their datasets—it may be unavoidable that these are exactly the type of data we need. Unfortunately, this kind of research is rare in the literature, probably because such data, being longitudinal in nature, are difficult to obtain and bring with them all sorts of practical problems. Assuming that longitudinal data are available, however, they can be used to answer some otherwise intractable questions. For example, what happens when lead examiners retire and different lead examiners guide training and operational rating? And how is the quality of rating affected when systems move from on-site to remote rating or hybrid system? Furthermore, do the changes in the composition of examinees in relation to, for example, their background, such as the first language (L1) or socioeconomic status, also influence the ratings (Bachman, 1990; McNamara, 1996)?

This is the first article in a two-part series on rater behaviour. It explores rater severity and consistency with longitudinal data collected between 2009 and 2019 in the speaking test of L2 Finnish in the National Certificates of Language Proficiency (NCLP) examination in Finland. This study is part of a larger project which approaches language testing as a dynamic social action. We investigate whether severity remains stable or fluctuates over time. More specifically, in the high-stakes context of the NCLP, we aim to answer two fundamental research questions:

1. Does rater severity change over a long period of time?
2. Does rater consistency change over a long period of time?

If rater severity or consistency changes measurably over time (i.e., beyond error of measurement levels), we will examine a third researchquestion:

3. Can the change(s) be attributed to major changes in the rating system, such as a change of lead examiner, the modus of training and rating (on-site or remote), or composition of the rater group?

Although other researchers have investigated the stability of rating characteristics over time, each using datasets with different characteristics (e.g., Lim, 2011; Myford & Wolfe, 2009), our research is unique for several reasons. We investigate rating stability over many years, but have multiple rating sessions per year. Thus, we have a richer dataset and can measure change with a finer resolution compared to recent research efforts (e.g., Lamprianou et al., 2021). Our dataset also includes information about structural changes, such as changes of the lead examiners, the impact of which has, to the best of our knowledge, never been investigated before. We use standard psychometric methods to analyse the data, as suggested in the literature (Lim, 2011; Myford & Wolfe, 2009).

Background

The NCLP is a legislation-based (Act 964/2004; Decrees 1163/2004 and 1109/2011) language examination system that is independent of any syllabus or curriculum. The exam assesses adults' command of nine languages at three different test levels (A1–A2, B1–B2, C1–C2). In this paper, the focus is on the intermediate level test of Finnish as an L2 (B1–B2), which is the biggest test in terms of the number of examinees. The test is designed to measure everyday language used for communication and functional purposes (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980). The assessment is guided by the descriptions of language proficiency in the Common European Framework of Reference for Languages (CEFR; Council of Europe (2001, 2020). The Framework of Finnish National Certificates (2011) and NCLP test specifications (2011), based largely on the CEFR, provide the guidelines for test content and item writing. The intermediate test is mostly taken by migrants who need to demonstrate their proficiency (B1 or higher) when applying for citizenship. According to examinee surveys between

2012 and 2021 (N=45,087), 79% of the examinees in the NCLP test use the certificate to apply for citizenship.

The current study focuses on the assessment in the speaking test of the Finnish intermediate level (B1–B2) examination. The speaking test is taken in a lab and it consists of four tasks (description, simulated conversations, reacting to short speaking situations, and opinion) measuring the language use in different situations. Performances are recorded for rating.

As a prerequisite, all raters need to successfully complete a specialized training course to be certified. After a successful completion of the certification course, they can register officially as raters and participate in further training and operational rating. Every rating session starts with a mandatory training session led by a lead examiner, after which performances are rated by comparing performances to NCLP criteria, linked to the CEFR. Examinees receive a separate score from each task, but the final proficiency level combines all four task-specific ratings.

From 2009 to 2019, the speaking performances were rated either in on-site rating sessions or remotely at home. On-site sessions start with face-to-face training, during which the raters listen to, assess, and discuss benchmark performances. The raters start to assess the performances immediately after the training. All performances are rated during two intensive 8-hour days. In on-site sessions, the raters work in the same place, and during breaks can talk with each other and with the lead examiner. In remote assessment, raters assess a set of performances before training and provide written justifications for their ratings. Remote assessment starts with an online training session led by a lead examiner using the video platform Zoom. In addition to live discussions, raters can express their opinions through chat. There are usually three separate training sessions within the first 2 days of the rating period and each rater must attend one of them. After training, rating is done remotely within a week. The remote rating is fragmented, as it is spread over several days. Although the lead examiner can be reached via email and by phone, communication with colleagues during rating is not possible, or at least is not supported by the system.

The lead examiners, experts in both Finnish and language assessment, have a prominent role in the assessment process. Besides leading the rater training sessions, they are responsible for the design of the examination, from test construction to marking and grading, including the item-writing process. They also select the benchmark samples used in the training. In addition to strengthening the use of assessment criteria, lead examiners set clear objectives for task fulfilment. These presentations of the lead examiner have an impact on rating. In case of disagreement on, for example, task fulfilment or proficiency level, the lead examiner makes the final decision. Lead examiners also give individualized feedback on operational ratings to the raters whose rating behaviour is found to be problematic. Unfortunately, due to the intense nature of rating (two 8-hour days for on-site rating), it is not practical to give feedback during the operational rating. Typically, testing organizations use procedures such as “backreading” (i.e., senior raters check the ratings of other raters for accuracy; see Wind, 2022) or “seed” responses (i.e., senior raters rate selected examinee responses and forward them to ordinary raters for quality control purposes; see Tisi et al., 2013). However, these are impractical to implement within the tight rating schedule in our context.

Over these 10 years (with 45 different rating sessions; around 4 per year), there were some structural changes which may have affected ratings. First, between 2009 and 2016, all training sessions and rating sessions were conducted as on-site face-to-face events. However, in the last 3 years of the research period (2017–2019), more than half of the rating and training sessions were conducted remotely. Second, three different lead examiners were responsible for rater training. In addition, there has been a large increase in the number of examinees along with a decline in test results (i.e., a higher proportion of examinees failed the test).

Literature review: Communal factors in rater severity and consistency

Recent research has described the multitude of factors influencing the quality of rating. Among other things and based on Knoch et al.'s theoretical model (2021), these factors include the raters and their training (in our context conducted by the lead examiners), rater experience, the rating context (including practices and conventions as well as the modus operandi, e.g., on-site/online), as well as the community of practice (CoP). The concept of the CoP has a prominent role in the literature and has been adapted by researchers such as Willey and Gardner (2011) and Herbert et al. (2014). The concept was introduced by the educational theorists Lave and Wenger (1991) to describe the process of learning taking place in a community. Therefore, the rating process is not only about evaluating learner performances through the lens of the rating criteria, but also about monitoring one's own ratings and those of others (Ahola, 2016, 2022). According to Lim (2011) and Davis (2016), overly severe, lenient, or inconsistent novice raters become more like their more experienced peers after a few scoring sessions. The community thus affects rater behaviour directly, but the influence might be short term (Lumley & McNamara, 1995).

A CoP has many ways to affect the behaviour of its members. In addition to informal learning (e.g., picking habits and practices through osmosis), communities often expect their members to undergo formal training. Rater training is generally considered to have a desirable effect on rater behaviour (e.g., Elder et al., 2005; Fulcher, 2003). It has been seen to contribute to both intra-rater reliability (Davis, 2016; Weigle, 1998) and inter-rater reliability (Fahim & Bijani, 2011; McIntyre, 1993; Shohamy et al., 1992; Weigle, 1994), although sometimes the effects were found to be small or short-lived (Bonk & Ockey, 2003; Knoch, 2011; Lumley & McNamara, 1995). Recent research (Attali, 2016; Davis, 2016; Rethinasamy, 2021) has shown that training had a slightly positive effect on the quality of rating overall. Official training, like standardization meetings, might also function as a shared experience enabling individual raters to co-evolve (Lamprianou et al., 2023). For example, training may help raters reconcile the way they interpret and use scales. Weigle (1994) provided evidence that training increased agreement among raters by clarifying the rating criteria.

Drawing on Knoch et al.'s (2021) theoretical model discussed above, we identified the modus of training and rating as an important factor affecting rating. The modus of training and rating can take three modes: on-site (face-to-face), online (regardless of specific time and location), or hybrid (combination of face-to-face and online). In recent

years, several tests have shifted to online training and rating due to considerations of accessibility and cost saving spurred by the COVID-19 pandemic (e.g., Isbell & Kremmel, 2020), although, to our knowledge, no relevant research from this period was available at the time we drafted this article. Knoch et al. (2018) compared the effectiveness of online training supported by a trainer and face-to-face rater training for novice writing raters. The results showed that there were no significant differences between the rating behaviour of these two groups. In the same context, Knoch et al. (2016) compared the effectiveness of training mode for novice speaking raters and reported that the face-to-face raters rated somewhat more consistently as a group compared to online raters. Based on these findings they suggested that more training or support is needed for online speaking raters.

More research is needed to examine the impact of online rater training in the context of speaking assessment. Based on previous research findings, online rater training can be as effective as face-to-face training in the context of writing assessment (e.g., Knoch et al., 2007; Wolfe et al., 2010), but these results cannot be directly transferred to speaking assessment due to the different nature of skill.

Data and methods

The study focuses on the speaking test of the Finnish Intermediate Level test in the NCLP and is based on the ratings of 94% of the examinees who took the test between 2009 and 2019. The study covers 45 different rating sessions (4–6 sessions per year), 104 raters, 59,899 examinees, and 175 different tasks. The excluded 6% of the examinees were rated by raters who had only a limited number of ratings per test (<10 examinees) or less than three rating sessions. On average, there were 1331 examinees per test, ranging from 402 to 2006. In addition, information about the lead examiners and the modus of rating during study period is included in the data.

Analysis and models

We analysed the rating data for all years in a single Many-Facets Rasch model (MFRM) using Facets software (Linacre, 1989, 2020). To investigate changes in rater severity and (in)consistency, we used generalized linear mixed models (GLMMs; Agresti, 2013; Bates et al., 2015). We used information about the lead examiner and the modus of rating (on-site/remote) and composition of the rater group as independent variables. The models were fitted with lmer and glmer functions from the lme4 package (Bates et al., 2015) on the R platform (R Core Team, 2020).

Requirement for data connectivity

To get a unique consistency and severity measure for each rater at each time point and to be able to compare a rater's measures at a given time point to measures at any other time point, we used the MFRM (Eckes, 2011; Myford & Wolfe, 2003, 2004). However, to do so, it is necessary to satisfy the connectivity requirement; in other words, it was necessary to identify common elements for all but one of the facets of measurement across the

45 rating sessions. In our dataset, there were three facets of measurement (i.e., the raters, the examines, and the items), so it was necessary to identify common elements for two of the three facets to verify that there were no "disjointed subsets" (Lim, 2011; Linacre, 1989). Within each exam year, our datasets are strongly connected by design.

To achieve data connectivity, we followed the standard procedure described in the literature (Lamprianou et al., 2021; Lim, 2009, 2011; Myford & Wolfe, 2009). First, we identified all the examinees who repeated the test during the period of 2009 to 2019. From those, we selected only the examinees whose "overall language skills" seemed to stay stable across examinations. To do so, we compared their test results on speaking, writing, reading, and listening between different tests. Every examinee whose performance/overall results did not change was considered a suitable person to form a link ("linking examinee") over rating sessions. In total, 1065 suitable examinees were identified and used to link the data. These examinees had taken 2650 tests and they were rated 3291 times.

It is reasonable to assume that achieving data connectivity through linking examinees, using the procedure described above could potentially affect the results of the MFRM. To investigate the robustness of our findings, we replicated the analysis using different subsets of linking examinees, but the results remained practically the same. The correlations between the rater severities of different analyses were around 0.97, suggesting that different subsets of linking examinees produced the same results for all practical intents and purposes.

Raters and examinees

The longitudinal data consisted of 104 trained (i.e., certified) raters who had a university degree in Finnish and who mostly teach Finnish as a first or second language but choose to work as raters as an extramural activity. All raters had completed a specific training course (as a certification step), approved by the National Agency for Education, which is a prerequisite for becoming an official NCLP rater. Almost all the raters had Finnish as their L1 and 94% of them were females (M_{age} = 51 years, SD = 10; range: 27–70.

Of the 59,899 examinees, 51% were females and 49% males. The age of 90% of examinees ranged from 20 to 50 years. They came from 185 countries, with the largest groups from Russia (19%), Iraq (9%), Estonia (6%), Turkey (4%), and Somalia (3%).

Lead examiners, modus of rating, and composition of the rater group as explanatory variables

As shown in Table 1, in the 10-year period, the L2 Finnish examination had three different lead examiners: Lead Examiner A covers 18 test rounds, Tests 1–18 (years 2009–2013), Lead Examiner B covers 8 test rounds, Tests 19–26 (years 2014–2015), and Lead Examiner C covers 19 test rounds, Tests 27–45 (years 2016–2019).

The modus of training and rating is a categorical variable with two classes: on-site and remote. Under the regime of Lead Examiners A and B, all sessions were conducted on-site. For Lead Examiner C, 9 remote and 10 on-site training and rating sessions were conducted, and all 9 remote sessions took place in the last 3 years of the research period.

Table 1. Test rounds and years covered by three different lead examiners.

Year	Tests rounds (remote)	Lead examiner
2009	1–2	A
2010	3–6	A
2011	7–10	A
2012	11–14	A
2013	15–18	A
2014	19–22	B
2015	23–26	B
2016	27–30	C
2017	31–36/35–36	C
2018	37–42/37, 39–42	C
2019	43–45/43, 45	C

The total number of raters working within each rating session ranged from 11 to 45 ($M = 32$; $SD = 8$). Forty or more raters took part in 18% of rating sessions, with 30 to 39 raters in 47% of rating sessions, and fewer than 20 raters in three sessions (7%). To study the effect of the cohesion of rater community on rater behaviour, we calculated the proportion of common raters (i.e., returning raters) for consecutive rating sessions (referred to as composition of the rater group). The proportion of common raters for consecutive rating sessions ranged from 17% to 78%, $M = 47\%$, $SD = 15\%$ (Figure 1). Starting from 2017 (Exam 31), the yearly number of examinations rose from four to six (see Table 1). As a result, the number of common raters for consecutive rating sessions decreased because not all raters had time to participate in all rating rounds and also because the raters were rotated between examinations and skills to assess.

Rating characteristics (dependent variables)

According to Linacre (2020), unreliable rating can be deduced from the rater's fit statistics produced by the MFRM model. There are no fixed cut-off values for fit statistics, but according to Wright and Linacre (1994), they depend on the assessment purpose. For our study, the acceptable values for infit mean square (MNSQ) and outfit MNSQ indices were set strictly to be below the threshold of 1.2; thus, we created a binary variable called misfit, where a value of 1 indicates inconsistency ($\geqslant 1.2$) and a value of 0 indicates consistency (< 1.2). Although different researchers sometimes use slightly different thresholds for the Rasch misfit (e.g., see Huang & Chen, 2022, for a less rigorous threshold), studies with similar aims to our study have used the same rigorous thresholds of around 1.2 (e.g., see Lamprianou et al., 2021).

Rater severity was measured in Rasch logits. The lower the logit measure, the more severe the raters. The mean rater severity was set to zero.

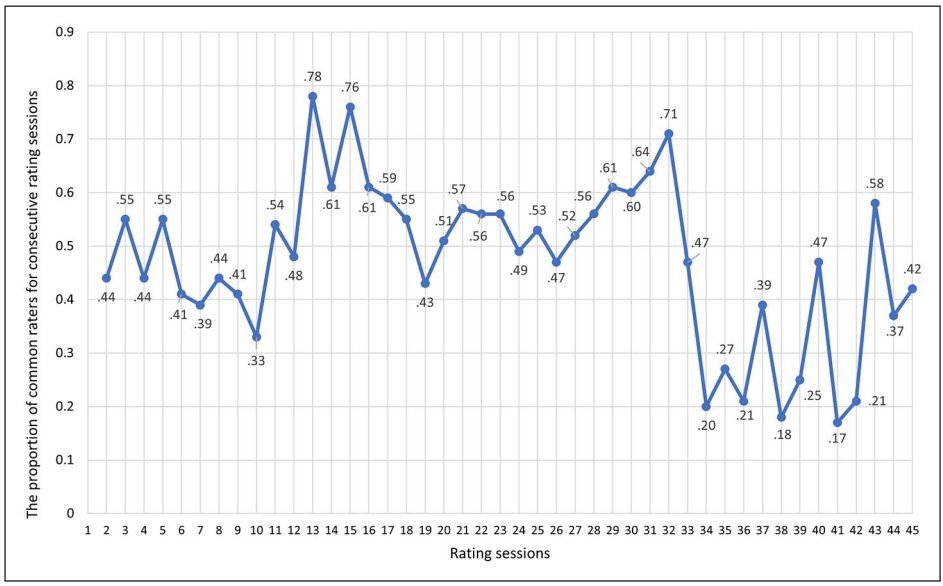

Figure 1. The proportion of common raters for consecutive rating sessions.

Table 2. Summary Rasch statistics for the three facets of measurement (examinees, raters, and items).

Statistics	Examinees	Raters	Items
Mean measure	−1.05 (0.88)	0.00[a] (0.12)	0.00[a] (0.05)
SD measure	3.12 (0.40)	0.78 (0.02)	0.85 (0.04)
Min measure	−8.64 (1.88)	−2.47 (0.13)	−1.61 (0.10)
Max measure	9.17 (1.88)	2.58 (0.13)	2.64 (0.07)
Adjusted (true) *SD*	2.97 (2.45)[b]	0.77	0.84
Homogeneity index	693,277.4 (df 59,898)***	65,424.5 (df 1417)***	71,383.8 (df 174)***
Separation	3.08	6.48	13.43
Strata	4.45	8.98	18.24
Reliability	0.90	0.98	0.99
Mean infit MNSQ	0.98	1.00	0.99
SD infit MNSQ	0.49	0.16	0.17
Mean outfit MNSQ	0.98	1.00	0.98
SD outfit MNSQ	0.62	0.19	0.19
N	59,899	1418	175

Note: The standard errors are given in parentheses. Examinee results with extremes shown in the table.
[a]The rater and item facets were set to have a mean measure of zero.
[b]Examinee results without extremes.
***$p = .001$.

Results

Summary Rasch statistics are presented in Table 2. The mean examinee ability was −1.05 logits (*SD* = 3.12; *SEM* = 0.88). The range of the examinee ability was very wide, ranging from −8.64 for the less able to 9.17 for the more able. The mean infit and outfit MNSQ values were 0.98.

The rater facet was constrained to have a mean rater severity of zero (*SD* = 0.78). Rater severity ranged from −2.47 to 2.58 logits, with smaller values indicating more severe raters. The standard errors of rater severities were for the most part small (from 0.07 to 0.14), indicating high measurement precision. The mean infit and outfit MNSQ values were 1.00 with standard deviations of 0.16 and 0.19, respectively.

The mean difficulty of the items (tasks) was also set to zero with a standard deviation of 0.85. The easiest item was −1.61 logits and the most difficult 2.64. The mean infit and outfit MNSQ values were 0.99 with standard deviations of 0.17 and 0.19, respectively.

The separation statistics (*separation, strata, reliability*) indicating the reproducibility of the measures were all high. The reliability for examinee measures was 0.90, suggesting that the test can differentiate between different examinees' proficiencies. The strata value denotes that our measurement system can separate about 4.5 statistically distinct proficiency levels. This is satisfactory for all practical intent and purposes, as the test aims to differentiate examinees on only three levels: "below Level 3," "Level 3," and "Level 4." The reliability for items was 0.99 and strata 18.24, also denoting significant differences among items in terms of item difficulty. It should be noted that high separation values can be obtained when the number of observations is large, which is the case here.

The reliability for rater measures was as high as 0.98 and the strata value was 8.98. These values suggest that our raters differ considerably terms of severity; our measurement system produced almost nine statistically distinct classes of rater severity.

The Rasch fit statistics for examinees, raters, and items are shown in Table 3. Using the rigorous cut-off value of 1.2 for the outfit MNSQ, as discussed above, we identified 14% of the raters and 11% of the items as misfitting. Using a moderate cut-off value of 1.3, we identified 6.4% of the raters and 4.6% of the items as misfitting the model. Using a cut-off value of 1.5, we identified 1.4% of the raters and 1.7% of the items as misfitting the Rasch model. Overall, we judge the data-model fit to be satisfactory for all practical intents and purposes of the study.

We also assessed overall model fit by a residual analysis. According to Linacre (2020), a satisfactory model fit is denoted when about 5% of the standardized residuals are outside ±2 and about 1% or less of the standardized residuals outside ±3. In our data, these figures were 5.3% and 0.6%. In addition, we explored the residuals related to particularly unmodeled noise and model underfit and found 2.8% of standardized residuals >2 and only 0.3% of standardized residuals >3. This supported our conclusion of adequate model fit to the data our purposes.

Rater severity changes over time

Figure 2 shows the distribution of rater severity over time. The *y*-axis represents the Rasch severity measure (in Rasch logits) and the *x*-axis exam order (ranging from 1 to

Table 3. Rasch fit statistics.

Statistics	Examinees	Raters	Items
Infit MNSQ ⩾1.2	12,114	150	19
Outfit MNSQ ⩾1.2	13,572	195	20
Observations	59,899	1418	175

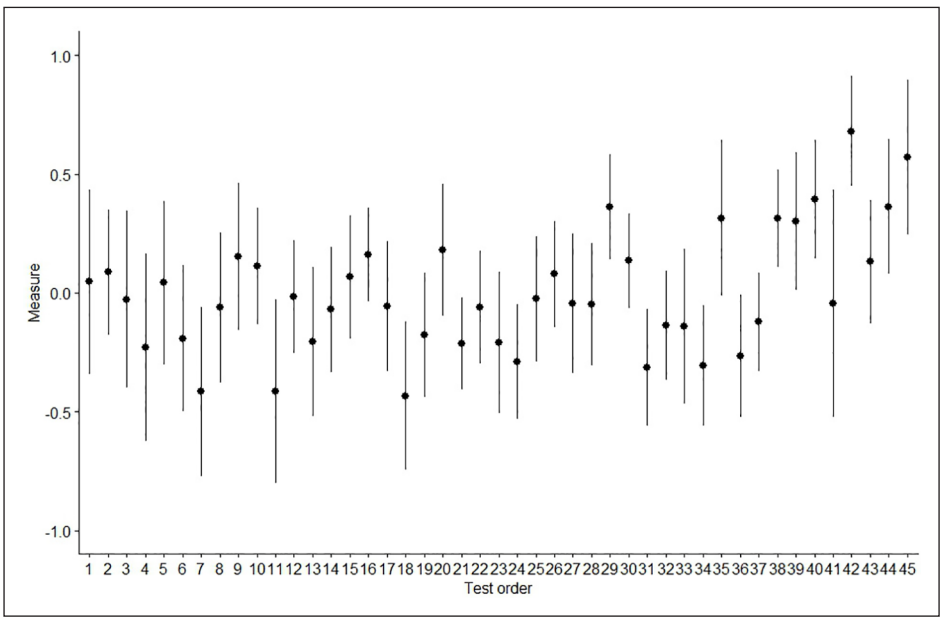

Figure 2. Mean rater severity with 95% confidence intervals. The y-axis represents the severity measure (logits) and the x-axis test order.

45). The average rater severity for some examinations seems to be much higher (or much lower) compared to others. The figure also illustrates that the mean rater severity measures tend to increase slightly towards the more recent tests, that is, raters (as a group) tend to become slightly more lenient towards the end of the period under study.

We used a GLMM, with raters as random effects, to explain rater severity by exam order. Table 4 shows that the raters become more lenient over time but only by a mere 0.01 logit per rating session, which is practically negligible. However, he intraclass correlation coefficient (ICC) of 0.29 suggests that, the same raters' severity measures yield a relatively low correlation (an explicit indication of their unstable severity on repeated occasions of rating sessions). Therefore, the passing of time alone (i.e., exam order) does not seem to be a major driver of rater severity. There is something else at play.

Our results showed that rater severity was unstable and that the raters became more lenient towards the more recent exam rounds. Below, we show how changes in structural factors in the rating system affected rater severity and its changes. First, we investigated

Table 4. The results of LME model explaining the rater severity by exam order.

		Measure		
Predictors		Estimates	95% CI	*p*
(Intercept)		−0.23	−0.35 to −0.11	<.001
Exam order		0.01	0.01 to 0.01	<.001
Random effects				
σ^2	0.45			
$\tau_{00\ rater}$	0.18			
ICC	0.29			
N_{rater}	104			
Observations	1418			
Marginal R^2 or conditional R^2	0.023/0.306			

Note: LME: linear mixed effects; CI: confidence interval; ICC: intraclass correlation coefficient.

if these changes could be explained by the change of the lead examiner. We then studied the modus of rating, on-site or remote, and the composition of rater group.

Lead examiner effect on rater severity

We used a GLMM to explain the rater severity using the three different lead examiner regimes as a categorical explanatory variable. The variable was added in the model as two dummy variables ("Lead Examiner B" and "Lead Examiner C") with "Lead Examiner A" as the reference category. The results are presented in Table 5. The estimate of "Lead Examiner C" was 0.24 and it was statistically significant at the .001 level. The coefficient of 0.24 is large and almost one third of the standard deviation of the distribution of rater severity (see Table 2). As a result, we deduce that under the regime of "Lead Examiner C," mean rater leniency seems to increase considerably.

Remote rating and training affects rater severity

We used a GLMM to explain rater severity using the modus of rating (on-site vs remote) as a categorical explanatory variable. The variable was added in the model as a dummy variable with "on-site rating" as the reference category. The results are presented in Table 6. The estimate of "remote rating" was 0.26 and was statistically significant at the .001 level. The coefficient of 0.26 is large and one third of the standard deviation of the distribution of rater severities (see Table 2). As a result, we conclude that the raters tended to rate more leniently (the mean rater severity decreases considerably) during the remote rating than on-site rating.

However, Lead Examiner C and remote rating are strongly intertwined, because all nine remote ratings and related trainings were conducted under the regime of Lead Examiner C. To find out whether the tendency towards leniency is also due to the modus

Table 5. The results of LME model explaining the rater severity by the regimes of three different lead examiners.

		Measure		
Predictors		Estimates	95% CI	*p*
(Intercept)		−0.11	−0.21 to −0.00	.043*
Lead examiner (B vs A)		0.03	−0.07 to 0.13	.581
Lead examiner (C vs A)		0.24	0.16 to 0.33	<.001***
Random effects				
σ^2	0.45			
$\tau_{00\ rater}$	0.19			
ICC	0.29			
N_{rater}	104			
Observations	1418			
Marginal R^2 or conditional R^2	0.021/0.309			

Note: LME: linear mixed effects; CI: confidence interval; ICC: intraclass correlation coefficient.
*$p < .05$; **$p < .01$; ***$p < .001$.

Table 6. The results of LME model explaining the rater severity by the modus of rating.

		Measure		
Predictors		Estimates	95% CI	*p*
(Intercept)		−0.05	−0.14 to 0.05	.320
Modus of rating (on-site vs remote)		0.26	0.16 to 0.37	<.001***
Random effects				
σ^2	0.45			
$\tau_{00\ rater}$	0.17			
ICC	0.28			
N_{rater}	104			
Observations	1418			
Marginal R^2 or conditional R^2	0.016/0.290			

Note: LME: linear mixed effects; CI: confidence interval; ICC: intraclass correlation coefficient.
*$p < .05$; **$p < .01$; ***$p < .001$.

of rating, that is, "remote rating," and not only due to Lead Examiner C, we analysed the data separately for the period of Lead Examiner C. Table 7 presents the results of the model predicting rater severity by modus of rating under the regime of Lead Examiner C only. The estimate of "remote rating" was 0.16 and was statistically significant at the .007 level. The coefficient of 0.16 is one fifth of the standard deviation of the distribution of rater severities under the regime of Lead Examiner C. In conclusion, both Lead Examiner C and modus of rating affected rater severity.

Table 7. The results of the LME model explaining the rater severity by the modus of rating under the regime of Lead Examiner C.

	Measure		
Predictors	Estimates	95% CI	*p*
(Intercept)	0.03	−0.08 to 0.15	.584
Modus of rating (on-site vs remote)	0.16	0.04 to 0.28	.007**
Random effects			
σ^2	0.42		
$\tau_{00\ rater}$	0.18		
ICC	0.30		
N_{rater}	95		
Observations	611		
Marginal R^2 or conditional R^2	0.010/0.311		

Note: LME: linear mixed effects; CI: confidence interval; ICC: intraclass correlation coefficient.
$^*p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Table 8. The results of LME model explaining the rater severity by the proportion of common raters for consecutive rating sessions.

	Measure		
Predictors	Estimates	95% CI	*p*
(Intercept)	0.20	0.05 to 0.36	.009
The composition of raters	−0.43	−0.69 to −0.17	.001
Random effects			
σ^2	0.46		
$\tau_{00\ rater}$	0.18		
ICC	0.28		
N_{rater}	104		
Observations	1314		
Marginal R^2 or conditional R^2	0.006/0.283		

Note: LME: linear mixed effects; CI: confidence interval; ICC: intraclass correlation coefficient.

Rater severity can be explained by the composition of raters

A GLMM was used to study whether rater severity can be explained by the composition of the group of raters (i.e., the proportion of returning raters for consecutive rating sessions), which is a numeric variable. The raters were modelled as random effects. The composition of the rater group seemed to predict rater severity (Table 8). Its coefficient of −0.43 was negative and statistically significant ($p = .001$), indicating that the raters became more lenient when they were rating in the group that had fewer raters from the previous rating session.

It should be noted that the proportions of common raters for consecutive rating sessions and the remote ratings are intertwined: two additional exam rounds (added from 2017, Exam 31) led to a drop in the proportion of raters invited to rate in consecutive tests and at the same time the remote ratings were introduced. Thus, it is not unexpected that both the remote rating and the rater group that had fewer raters from the previous test round affect rater severity in the same way (towards leniency).

Rater consistency is not affected by lead examiner, modus of rating and training, or composition of raters

We used a GLMM, with raters as random effects, to explain rater (in)consistency by exam order, lead examiner, modus of rating, and composition of raters. For each of the explanatory variables in the model, we estimated a coefficient (i.e., "odds ratio") showing the degree to which the variable increased (or decreased) the odds of a rater to be identified as being inconsistent. The odds ratio of passing of time (odds ratio 1.01, $p = .467$), the odds ratios of Lead Examiners B and C (Lead Examiner B: odds ratio 0.91, $p = .719$; Lead Examiner C: odds ratio 0.86, $p = .465$), and the odds ratio of "remote rating" (odds ratio 1.18, $p = .483$) were all near 1, except the odds ratio of "composition of group of raters" (odds ratio 1.91, $p = .299$), and none was statistically significant. This suggests that none of the variables influenced (beyond randomness) the chances of a rater to be identified as being inconsistent. The ICCs of 0.37 denoted that raters have moderate within-rater correlation of odds to be classified as inconsistent, suggesting both that inconsistency is more like a random phenomenon (i.e., raters occasionally behave in an inconsistent manner in their ratings), and that inconsistency accumulates (more likely) only for specific raters. In other words, inconsistency may, to some degree, be a personal characteristic, rather than the effect of the variables included in the model. However, the results highlighted that raters behaved for the most part in a consistent manner in their ratings across the study period.

Discussion

This study examined rater consistency and severity changes over 10 years (2009–2019) in a Finnish as a second- or foreign-language high-stakes examination. As explanatory factors for the changes, three variables were studied: changes of lead examiner, the modus of training and rating (on-site or remote), and the composition of the group of raters (i.e., the proportion of returning raters in consecutive tests). These three variables are important components of the model proposed by Knoch et al. (2021) regarding the factors affecting the quality of rating.

Earlier studies, such as Myford and Wolfe (2009) and Lamprianou et al. (2021), demonstrated that rater severity effects are not stable over time. Congdon and McQueen (2000) found that raters have different trends in their scoring over time. Our longitudinal study confirms the earlier results that there are changes in rater severity over time. First, raters as a group became slightly more lenient towards the end of the studied period. Second, based on the ICC, the raters' severity measures fluctuated over time, hence, were

unstable. When combining these findings, we found that the passage of time alone did not explain the change in rater severity. Moreover, rater consistency was not affected by passing of time. The inconsistent rating behaviour was occasional, although the ICC of 0.37 suggested that some raters tend to behave in a more inconsistent manner in their ratings compared to other raters.

Rater severity has a direct impact on ratings, meaning that fluctuations in rater severity are a threat to the validity and reliability of the ratings. That is why most test systems have procedures for controlling (e.g., rater training) and monitoring (e.g., statistical analyses and feedback) rater behaviour. Usually, these procedures are applied on an exam-by-exam basis or within each exam marking period. Individual rater severity and consistency fluctuations in consecutive rating sessions and during the session can be detected more easily than a gradual change in the severity/leniency of a whole rater group. To explore a long-term trend of rater severity, the test system must put effort into data collection, storage, and analysis. In addition, in the case of discovering a change, it is crucial to examine its causes. What factors make raters more lenient or severe?

What turned out to have an impact on the general rater severity was the lead examiner. Under the regime of Lead Examiner C, the raters as a group became more lenient compared to the regimes of Lead Examiners A and B. This is in line with the model of Knoch et al. (2021), which suggests that training and feedback are an important factor affecting ratings. In the context of our study, the lead examiners lead the training, have the main role in decisions on task fulfilment, choose the benchmarks for the rating, and lead discussions related to the assessment criteria. Lead examiners also give individualized feedback to the raters whose rating behaviour is found to be problematic. The feedback covers the statistics of the ratings and any that performances received biased ratings. Previous research findings on the impact of feedback are contradictory. For example, Hoskens and Wilson (2001) argued that the feedback caused a drift towards the mean, but Knoch (2011) argued that raters were unsuccessful in incorporating the feedback into their ratings. Based on the experience in NCLP, it seems that the rating behaviour gets better right after feedback, especially with raters with a positive attitude towards feedback, but the effect is short term.

In conclusion, lead examiners have a considerable impact on the raters' general rating policy. Differences in rater severity between lead examiners may stem from their view of language and what they value in speaking and may contribute to the rating policy being implemented in rater training. Therefore, paying attention to how lead examiners conduct rater training and monitoring and examining the quality of ratings of the lead examiners are crucial.

Even as the lead examiner influenced rater severity as a whole, the differences in the range of rater severities across rating sessions remained stable. The lead examiner can thus affect the line of rating to some extent but cannot eliminate the difference between the most severe and most lenient rater, which has also been shown in previous studies (e.g., Fahim & Bijani, 2011; Shohamy et al., 1992).

If severity really changes with the lead examiner, it would be problematic for the reliability and validity of the test system, as it would suggest that general rater severity would vary along one person's individual line of rating. We suggest, however, that while the lead examiner certainly has an effect in the rater community, the change in the

general line of rating has more to do with the modus of rating and rater community than individual differences.

We found that the raters tended to rate more leniently when the rating and training were conducted remotely rather than on-site. Again, this is in line with the model by Knoch et al. (2021), who identified the rating conditions as important factors affecting ratings. Remote training and ratings were mostly conducted at the end of the studied 10-year period, all under the regime of Lead Examiner C. This result aligns with our findings concerning the effect of the lead examiner. It was also shown that under the regime of Lead Examiner C, raters rated more leniently remotely than they did on-site.

Since the modus of rating seems to affect rater severity, it is important to consider potential reasons for this effect. The structure of the training is standardized and led by the lead examiner regardless of the modus of the training. However, there are clear differences between on-site and remote training related to the number of training sessions per exam round, the time of rating of the training samples, and the practical (technical) realization of the sessions, all of which affect interaction between the raters and the lead examiner.

There are usually three training sessions for one examination and rating round when remote rating is used and one session in on-site training. In practice, this affects the size of the training groups and, thus, the interaction and group dynamics directly. The good thing about on-site rating is that all raters take part in the same training, which means they hear the same arguments, guidelines, and policies. Based on experience, interaction between raters and lead examiner is superior in on-site to remote training. However, in a bigger group, it is also easier to opt out of the discussions and be a bystander. In on-site training, the lead examiner usually collects the assessments at the same time by hand vote. The influence of strong characters, so-called thought leaders actively expressing their own assessments and reasoning, becomes more obvious. The group dynamics that develop when trainers (who could be considered equal to the lead examiner) are available for live discussions can be effective in shaping rating behaviour (see Anderson et al., 2015). Live discussions can affect change due to persuasion through human interaction; as a result, self-access online training (without the interaction with the trainer) may be less effective and less efficient (time-wise) compared to training sessions mediated by rater trainers.

One factor that may affect rating is how soon after training the actual rating starts. Log data from remote ratings show that only a few raters start to work right after the remote training. Most of them start within 1 to 3 days. In on-site rating, raters start to work intensively right after training. Some previous studies (e.g., Lumley & McNamara, 1995) found that raters' accuracy improved immediately after training but that the effect did not last very long.

The informal and formal interactions between raters can also affect their ratings. These discussions can influence how raters understand and interpret the rating criteria and task fulfilment and how they weigh the different criteria. During on-site rating sessions, the raters work together and can interact during breaks. The affinity between the NCLP raters is high, and they can easily reflect their rating policy with other raters. Ahola (2016) has shown that among the raters, severity was a more desired feature than leniency.

It seems that the rater community and its practices affect the raters so that when surrounded by their colleagues and working intensively during the 2-day on-site rating sessions, the raters assess more harshly than when they are on their own, rating at home and their work is divided into shorter periods over several days. Our results align with Knoch et al.'s (2021) model, which identified the CoP as an important determinant of rating quality. Our results also showed that raters became more lenient when rating in a group which had fewer raters from the preceding test round. This aligns with the findings of previous research that a tighter community makes a rater more severe (Lamprianou et al., 2021).

Thus, based on the findings, to maintain the quality of ratings, test validity, reliability, and fairness, the test organization must carefully consider how to conduct training and ratings. Both remote and on-site ratings can be efficient from the rater consistency point of view.

In the current study, rater consistency, intra-rater reliability were not affected by the lead examiner, the composition of rater group, or the modus of rating and training. The rater training, which has been seen to contribute positively to intra-rater reliability in previous studies (e.g., Leckie & Baird, 2011; Lim, 2011), seems to be equally effective no matter how it is organized, whether remotely or on-site. In addition, raters can concentrate on their work at home so that a fragmented rating schedule, workload (a combination of "actual work" and rating), and other distractors at home (e.g., interruptions by family members) do not lead to inconsistent ratings more often than working on-site. The result also suggests that colleagues do not have the same effect on consistency as they do on severity.

Conclusion

In this article, we investigated factors that might affect raters' severity in the NCLP examination system in Finland. We discussed evidence that the CoP, encompassing the rater trainer (lead examiner), training methods, the rating process, and the composition of raters, influenced rater severity. However, it did not have an impact on rater consistency.

There are some potential factors that might be considered when reflecting oning the findings of our study. In real life, global events and subsequent processes such as migration naturally affect the composition of examinees (in an examination like ours which is used, e.g., in applying for citizenship). Many structural changes, such as a change of lead examiner, a change of modus of training and rating, and an increase in the number of tests, occurred simultaneously within a short time. It is challenging, if not impossible, to isolate one factor from another when interpreting the results.

Reflecting on our research design, a more in-depth approach using qualitative data could have helped to illuminate several of the issues discussed in this paper, such as the interaction between the lead examiner and the raters, or the impact of on-site/online modus on the formal and informal communication between the members of the CoP. However, these kind of qualitative data were not readily available to us. We would like to encourage the research community to invest more resources in mixed methods studies in this area of research, to extend their quantitative findings with insights gained by in-depth qualitative data.

Finally, a word of caution for practitioners. Our findings have shown that both the modus of training or rating and the change of the lead examiner may have a significant impact on ratings. In the real world, and as a result of external stimuli (e.g., the COVID-19 pandemic), testing organizations may be forced to make sudden changes in their mode of training or rating. In addition, key personnel such as the lead examiner, or senior raters, may retire or move to a new career. These changes may cause shocks to the CoP and could negatively affect the quality of rating. Testing organizations and assessment regulators must have appropriate procedures in place to ensure that major changes in the assessment ecosystem will not have a negative impact on the quality of rating and will not disturb the stability of the system.

Acknowledgements

The authors would like to thank Mia Halonen, Sari Ahola, Ari Huhta, Tuija Hirvelä, Sari Ohranen, and Riikka Ullakonoja for providing helpful comments and support throughout the process. They would also like to thank the anonymous reviewers for their comments during the peer review process.

Author contributions

Reeta Neittaanmäki: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Writing—original draft; Writing—review & editing.
Iasonas Lamprianou: Conceptualization; Formal analysis; Investigation; Methodology; Supervision; Writing—review & editing.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The first author (Neittaanmäki) is employed as a research statistician for the National Certificates of Language Proficiency (NCLP) examination system. However, she was not involved in planning and drafting tasks nor assessing performances. NCLP is administered by the Finnish National Agency for Education, funded by the Ministry of Education and Culture and operated by the University of Jyväskylä. NCLP researchers are complementarily financed by the University of Jyväskylä.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was conducted as part of the Broken Finnish project funded by the Research Council of Finland (grant no. 315581) and the University of Jyväskylä.

ORCID iDs

Reeta Neittaanmäki https://orcid.org/0000-0001-9741-5584
Iasonas Lamprianou https://orcid.org/0000-0001-7637-615X

References

Act on the National Certificates of Language Proficiency 964/2004. https://www.finlex.fi/fi/laki/ajantasa/2004/20040964
Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley & Sons.

Ahola, S. (2016). Puhetta arvioinnista: Yleisten kielitutkintojen arvioijien käsityksiä arvioinnista [Raters' beliefs and views of assessment in the National Certificates of Language Proficiency]. In A. Huhta & R. Hildén (Eds.), *Kielitaidon arviointitutkimus 2000-luvun Suomessa* [*Research on language assessment in 21st century Finland*] (pp. 89–109). Suomen soveltavan kielitieteen yhdistys. AFinLA-e: soveltavan kielitieteen tutkimuksia, 9. http://journal.fi/afinla/article/view/60848

Ahola, S. (2022). *Rimaa hipoen selviää tilanteesta—Yleisten kielitutkintojen suomen kielen arvioijien käsityksiä kielitaidon arvioinnista ja suullisesta kielitaidosta* [Barely passing the test task—NCLP Finnish raters' beliefs about language assessment and spoken language skills] [Doctoral dissertation, University of Jyväskylä]. JYX Digital Repository. http://urn.fi/URN:ISBN:978-951-39-9005-3

Anderson, D., Irvin, S., Alonzo, J., & Tindal, G. A. (2015). Gauging item alignment through online systems while controlling for rater effects. *Educational Measurement: Issues and Practice*, *34*(1), 22–33. https://doi.org/10.1111/emip.12038

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99–115. https://doi.org/10.1177/0265532215582283

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Barkaoui, K. (2010). Variability in ESL Essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

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

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa

Canale, M. A., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*(2), 163–178. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x

Council of Europe. (2001). *Common European Framework of Reference for Languages*: Learning, teaching, and assessment. https://rm.coe.int/1680459f97

Council of Europe. (2020). CEFR Companion Volume: Enhancing engagement in language education. http://www.coe.int/lang-cefr

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117–135. https://doi.org/10.1177/0265532215582282

Decree on the National Certificates of Language Proficiency 1109/2011. https://www.finlex.fi/fi/laki/alkup/2011/20111109

Decree on the National Certificates of Language Proficiency 1163/2004. https://www.finlex.fi/fi/laki/ajantasa/2004/20041163

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang. https://doi.org/10.3726/978-3-653-04844-5

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Fahim, M., & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing*, *1*(1), 1–16. https://www.ijlt.ir/article_114349.html

Fulcher, G. (2003). *Testing second language speaking* (1st ed.). Routledge. https://doi.org/10.4324/9781315837376

Herbert, I. P., Joyce, J., & Hassall, T. (2014). Assessment in higher education: The potential for a community of practice to improve inter-marker reliability. *Accounting Education*, *23*(6), 542–561. https://doi.org/10.1080/09639284.2014.974195

Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement*, *38*(2), 121–145. https://doi.org/10.1111/j.1745-3984.2001.tb01119.x

Huang, J., & Chen, G. (2022). Individualized feedback to raters in language assessment: Impacts on rater effects. *Assessing Writing*, *52*, 100623. https://doi.org/10.1016/j.asw.2022.100623

Isbell, D. R., & Kremmel, B. (2020). Test Review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, *28*(2), 179–200. https://doi.org/10.1177/0265532210384252

Knoch, U., Fairbairn, J., & Huisman, A. (2016). An evaluation of an online rater training program for the speaking and writing sub-tests of the Aptis test. *Papers in Language Testing and Assessment*, *5*(1), 90–106. https://doi.org/10.58379/xdyp1068

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox Publishing.

Knoch, U., Fairbairn, J., Myford, C., & Huisman, A. (2018). Evaluating the relative effectiveness of online and face-to-face training for new writing raters. *Papers in Language Testing and Assessment*, *7*(1), 61–86. https://doi.org/10.58379/zvmm4117

Knoch, U., Read, J., & Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Lamprianou, I. (2018). Investigation of rater effects using social network analysis and exponential random graph models. *Educational and Psychological Measurement*, *78*(3), 430–459. https://doi.org/10.1177/0013164416689696

Lamprianou, I., Tsagari, D., & Kyriakou, N. (2021). The longitudinal stability of rating characteristics in an EFL examination: Methodological and substantive considerations. *Language Testing*, *38*(2), 273–301. https://doi.org/10.1177/0265532220940960

Lamprianou, I., Tsagari, D., & Kyriakou, N. (2023). Experienced but detached from reality: Theorizing and operationalizing the relationship between experience and rater effects. *Assessing Writing*, *56*, 100713. https://doi.org/10.1016/j.asw.2023.100713

Lave, J., & Wenger, E. (1991). *Situated learning. Legitimate peripheral participation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815355

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*(4), 399–418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Lim, G. S. (2009). *Prompt and rater effects in second language writing performance assessment* [Unpublished doctoral dissertation]. University of Michigan.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. https://doi.org/10.1177/0265532211406422

Linacre, J. M. (1989). *Many-facet Rasch measurement* (2nd ed.). MESA Press.

Linacre, J. M. (2020). *A user's guide to FACETS: Rasch-model computer programs*. Winsteps.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions*, *13*(4), 425–444. https://doi.org/10.1177/016327879001300405

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples* [Unpublished master's thesis]. University of Melbourne.

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rethinasamy, S. (2021). The effects of different rater training procedures on ESL essay raters' rating accuracy. *Pertanika Journal of Social Sciences and Humanities*, *29*(Suppl. 3), 401–419. https://doi.org/10.47836/pjssh.29.s3.21

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*(1), 27–33. https://doi.org/10.2307/329895

Simpson, L., & Baird, J.-A. (2013). Perceptions of trust in public examinations. *Oxford Review of Education*, *39*(1), 17–35. https://doi.org/10.1080/03054985.2012.760264

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). *A review of literature on marking reliability research* (Report for Ofqual 13/5285). National Foundation for Educational Research.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223. https://doi.org/10.1177/026553229401100206

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Willey, K., & Gardner, A. (2011, 27–30 September). Building a community of practice to improve inter marker standardisation and consistency. In J. Bernardino & J. C. Quadrado (Eds.), *Proceedings of the SEFI 2011* (pp. 666–671), Lisbon, Portugal.

William, D. (1996). Standards in examinations: A matter of trust? *The Curriculum Journal*, *7*(3), 293–306. https://doi.org/10.1080/0958517960070303

Wind, S. A. (2022). Rater connections and the detection of bias in performance assessment. *Measurement: Interdisciplinary Research and Perspectives*, *20*(2), 91–106. https://doi.org/10.1080/15366367.2021.1942672

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment*, *10*(1), 1–21. https://ejournals.bc.edu/index.php/jtla/article/view/1601

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. https://www.rasch.org/rmt/rmt83b.htm