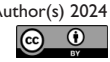
Original Manuscript

LANGUAGE TESTING

Language Testing
1–21



DOI: 10.1177/02655322241239362
journals.sagepub.com/home/ltj


All types of experience are equal, but some are more equal: The effect of different types of experience on rater severity and rater consistency

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 different types of rater experience over a long period of time. The article is based on longitudinal data collected from 2009 to 2019 from the second language Finnish speaking subtest in the National Certificates of Language Proficiency in Finland. The study investigated whether rater severity and consistency are affected differently by different types of rater experience and by skipping rating sessions. The data consisted of 45 rating sessions with 104 raters and 59,899 examinees and were analyzed using the Many-Facets Rasch model and generalized linear mixed models. The results showed that when the raters gained more rating experience, they became slightly more lenient, but different types of experience had quantitatively different magnitudes of impact. In addition, skipping rating sessions, and in that way disconnecting from the rater community, increased the likelihood of a rater to be inconsistent. Finally, we provide methodological recommendations for future research and consider implications for practice.

Keywords

Generalized linear mixed models, Many-Facets Rasch model, rater consistency, rater experience, rater severity, speaking assessment

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

Introduction

In every examination, human raters are perceived to be a potential source of error, which affects the validity and reliability of measurement. The pertinent literature has coined the term *rater effects* to identify a large set of idiosyncratic rating behaviors such as extreme leniency or severity, or a tendency to use only parts of a rating scale (Leckie & Baird, 2011; Wiseman, 2012; Wolfe, 2004). The term rater effects convey negative connotations; see, for example, the definition given by Wolfe and McVay (2012), who suggested that rater effects are "patterns of ratings that contain measurement errors" (p. 32). Typically, policymakers wish to minimize rater effects (i.e., errors) and one way to achieve this is by employing experienced raters. The pertinent literature has not yet satisfactorily answered some fundamental questions regarding what qualifies as rater experience and whether different types of experience shield raters from rater effects. In response to this pressing need, our research aims to address the "definitional cacophony" of rater experience (Lamprianou et al., 2023) by operationalizing different types of experience and investigating how they may influence rater effects differently.

Recent research (Lamprianou et al., 2023) has suggested a framework regarding the nature of rater experience as multifaceted, continuous, shared and temporal. Rater experience is multifaceted because it can be gained from different contexts and after using different scales. For example, it should not be taken as a given that experience and skills gained in one examination system are directly transferable to a different one (Knoch et al., 2020). Experience may be perceived as continuous and additive because the raters accumulate experience as they participate in more rating sessions. It is also shared because experience is realized as a by-product of being a member of a community of practice (CoP) (Lave & Wenger, 1991). Finally, experience is only temporal and may quickly become obsolete if a rater disconnects from his or her CoP.

Most studies on the importance of experience on rating behavior are related to the assessment of writing (e.g., Attali, 2016; Barkaoui, 2010b; Cumming, 1990; Erdosy, 2004; Lim, 2011; Lumley, 2005; Wolfe, 2005) rather than speaking performance. Mohd Noh and Mohd Matore (2022) argued that the relationship between experience and rating remains "underexplored" in speaking assessment. This may be the result of the relative scarcity of speaking components compared to writing components in language testing (Fan & Knoch, 2019), especially in the past, but this is a gap that nevertheless needs to be filled. Though speaking and writing assessments share some common elements, Davis (2012) emphasizes that research findings from writing assessments should not be blindly applied to the context of speaking assessments. This is because they differ not only in the language features to be evaluated, but also in the manner in which raters engage with examinee responses.

The wider literature linking rater effects and rater experience has largely been filled with contradictory findings, a fact that has not evaded the attention of the researcher community (Mohd Noh & Mohd Matore, 2022). The same problem has been observed in the subfield of speaking assessment, where the pertinent studies (Bonk & Ockey, 2003; Davis, 2016; Kim, 2011, 2015; Mohd Noh & Mohd Matore, 2022) have also produced contradictory findings. For example, Davis (2016) reported that experience had an insignificant effect on rater severity and consistency, whereas Kim (2015) found that

experience was an important variable affecting rating behavior in speaking assessment. It has been observed that raters with different experiences may interpret and apply the assessment criteria in different ways (Eckes, 2009; Khabbazbashi & Galaczi, 2020; Weigle, 2002). This holds true even if raters have experience in teaching learners from different backgrounds (Isaacs & Thomson, 2013; Winke et al., 2012). As the findings are contradictory and "experience" is defined differently in different studies, it is obvious that more research is needed to better understand which types of experience are relevant and how they affect rater behavior to ensure rating quality, especially in high-stakes tests.

This is the second article in a two-part series on rater behavior. The first article (Neittaanmäki & Lamprianou, 2024) focuses on how 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, affect rater effects. In this article, we use a framework for investigating how different operationalizations of experience affect rater effects (Lamprianou et al., 2023), that is, rater severity and consistency, in a high-stakes speaking examination. We are particularly interested to investigate the temporal nature of experience, by examining how absences from rating sessions, which means disconnecting from the CoP, may affect the manifestation of rater effects. As far as we know, this is the first study of its kind to focus exclusively on speaking assessment while using longitudinal data and multiple operationalizations of experience.

For the purposes of this study, we use longitudinal data (2009–2019) from the intermediate-level speaking test of second language (L2) Finnish in the National Certificates of Language Proficiency (NCLP) examination in Finland. We focus on the following questions:

Rater severity:

1a. Is rater severity affected differently by different types of rater experience?
1b. Is rater severity affected by missing rating opportunities, that is, by disconnection from the CoP?

Rater consistency:

1a. Is rater consistency affected differently by different types of rater experience?
1b. Is rater consistency affected by missing rating opportunities, that is, by disconnection from the CoP?

Lamprianou et al. (2021) investigated how skipping just one annual test affects the severity or the propensity of the raters to yield an aberrant rating pattern in the context of writing assessment. We apply here the same methodology but extend the study in three ways: we (a) investigate speaking rather than writing assessments, (b) use data with much higher granularity, and (c) use data from a different country and context to confirm the findings of past research.

The high granularity of our dataset is essential when it comes to investigating missing rating opportunities. The dataset used in this study allows researchers to take a micro perspective to investigate whether the extent of the gap (i.e., skipping different numbers of consecutive rating opportunities) is significant because there are many rating cycles per year. In other words, the data enable us to investigate the possible alienation process

where raters gradually lose their shared understanding with the CoP by accumulating gaps (i.e., by skipping more than one consecutive rating opportunity). It makes sense to assume that alienation is a gradual process and not an abrupt event caused by missing a single rating opportunity (cf. Lamprianou et al., 2021).

Context of the study

The study was conducted in Finland, in the context of the NCLP, and its intermediate level examination of Finnish (B1–B2). The examination levels are linked to the descriptions of language proficiency in the Common European Framework of Reference for Languages (Council of Europe, 2001, 2020). The intermediate test is a high-stakes test for adults that is used, for example, in demonstrating language skills for Finnish citizenship and applying for a job or a study place. According to a survey between 2012 and 2021, applying for citizenship was the most common and important reason for taking the test because 79% of participants ($N = 45{,}087$) reported using the certificate to apply for citizenship.

The study focuses on raters who assess speaking performances in the NCLP examination. These raters play a crucial role in the NCLP system, and their rating experience is the central aspect being examined. The regulation of raters in the NCLP system is governed by the pertinent acts and regulations (Act 964/2004; Decrees 1163/2004 and 1109/2011). These guidelines outline the qualifications required for raters, compulsory rater training, and regular participation in rating sessions. To ensure transparency and accountability, all raters are registered in the rater register maintained by the National Agency for Education.

The study specifically examines the impact of different types of rating experience on rater behavior in the assessment of intermediate level (B1–B2) speaking examinations. These examinations are conducted in a language lab, where performances are recorded for assessment purposes. The speaking subtest, lasting 20 minutes, consists of four distinct tasks that evaluate language proficiency in different situations and contexts. Prior to commencing the rating process for recorded performances, raters receive mandatory training either remotely or on-site from a lead examiner. During the training, raters listen to, assess, and discuss benchmark performances (see more about training in Neittaanmäki & Lamprianou, 2024). The speaking performances are rated using the NCLP criteria, which are linked to the CEFR.

Literature review: Rater experience in rater severity and consistency

The ratings assigned by raters may typically be influenced by multiple factors. Both the variability between raters and within raters can be influenced by specific rater characteristics, such as different types of experience, including their rating and teaching experience (Royal-Dawson & Baird, 2009). Several studies have explored the effect on ratings, when raters share a common language background with examinees or possess familiarity with the examinees' first language (L1) (e.g., Carey & Szocs, 2023; Huang et al., 2016;

Miao, 2023; Park, 2020; Winke et al., 2012). These studies have demonstrated that experience (i.e., familiarity) with the examinees' language background can facilitate rating, but this may also result in increased leniency. On the contrary, sometimes familiarity could potentially lead to the opposite effect (i.e., to increased severity), especially if the raters have negative experiences with speakers of the familiar L1. This phenomenon has been observed in several pertinent studies (Brennan & Brennan, 1981; Lindemann, 2005).

The relationship between rater experience and rater severity and consistency has been the focus of a growing number of studies. However, as stated above, most have been in the context of writing assessment (e.g., Ahmadi Shirazi, 2019; Barkaoui, 2010a, 2010b; Leckie & Baird, 2011; Lim, 2011; Şahan & Razi, 2020; Song & Caruso, 1996; Weigle, 1998, 1999) and their findings are not directly applicable to speaking assessment (Bonk & Ockey, 2003; Davis, 2016; Kim, 2011). The fact that these studies have operationalized experience differently may also make the comparability and applicability of their findings even more difficult.

For instance, Davis (2016) conducted a study to examine the impact of experience on the rating behavior of 20 experienced teachers who were otherwise inexperienced with that particular spoken language assessment. Results revealed that "posttraining experience" had limited influence on rater consistency and severity. However, it should be noted that this result might be attributed to the short data collection period of approximately 2 weeks. In another study, Bonk and Ockey (2003) analyzed group oral discussions among Japanese students in their L2 (English). The findings indicated that "returning" raters (with recent experience) were slightly more severe and consistent, but new raters were more inconsistent. Kim (2011) compared "novice," "developing," and "expert" raters and found that novice raters exhibited instability in severity compared to experts. In a somewhat similar study, Kim (2015) explored how novice, developing, and expert raters utilized an analytic scoring rubric in a speaking assessment across three rating sessions. The study demonstrated novice raters showed slow improvement in their rating performance, while experienced raters demonstrated stable rating patterns over time.

However, some results stemming from the writing assessment studies are possibly applicable to speaking assessment as well. In the context of writing assessment, Lim (2011) found that novice and experienced raters did not always demonstrate different severity and consistency. Leckie and Baird (2011) contrasted three groups of raters with different experience: team leaders, experienced raters, and new raters. They found that less and more experienced raters did not differ significantly in severity or consistency on essay scoring. Yet other studies have also yielded supportive results (Alp et al., 2017; Attali, 2016). In contrast, some research findings suggest that experience could make raters appear more lenient (e.g., Song & Caruso, 1996; Weigle, 1999). Weigle (1998) found that more experienced raters were less severe but more consistent. However, after training, consistency was improved significantly. Song and Caruso (1996) reported that experience was significantly related to leniency.

How raters interpret and use the rating criteria has been found to be one of the most important determinants of severity and rating variability (Barkaoui, 2010b; Lumley & McNamara, 1995; Weigle, 2002). Previous studies have shown that experienced raters

either use only certain individual analytical criteria (Eckes, 2008) or rely on factors not included in the criteria to feed into proficiency level judgments (Barkaoui, 2010a; Orr, 2002). For example, Eckes (2008) found that some raters tended to focus on different criteria compared to other raters. Eckes' results are also supported by Orr (2002), who indicated that, over time, raters considered factors that were not reflected in the assessment criteria and had difficulties adhering to the criteria. Other studies have indicated that experienced raters tend to assess holistically so they may pay less attention to rating criteria (Barkaoui, 2010b, 2011). In contrast, Cumming (1990) reported that experienced raters used a wider range of criteria compared to other raters. Ahmadi Shirazi (2019), on the other hand, found no statistically significant interaction between experience (novice or experienced rater) and rating scale (analytic or holistic scale).

Consistency in the interpretations of rating criteria is one of the biggest challenges in both rating and rater training. Even if raters understand the rating criteria similarly, they may value different features of the language in different ways and, thereby, may weight and apply different analytical criteria in their scoring (Ang-Aw & Chuen Meng Goh, 2011; Kim, 2011, 2015; Lumley, 2002; Orr, 2002; Şahan & Razi, 2020). Pollitt and Murray (1996) noted that raters' attention to different performance features also changes across examinees' proficiency levels.

Leaving aside the limited volume of published research in the field of speaking assessment, the main weakness of the existing literature is that researchers often categorize raters in obscure groups (e.g., "developing" or "experienced" raters) instead of employing more precise operational definitions. This may have led to the conflicting findings of similar research and has hindered the comparability of research outcomes. In the next section, we describe how we have operationalized experience in different ways to show that the type of experience may be a significant determinant of rater effects.

Data and methods

The data consist of two parts: (a) operational rating data from Finnish intermediate speaking subtests in NCLP over the last 10 years 2009–2019 covering 45 different examination sessions, 59,899 examinees, 175 tasks and 104 raters, and (b) background information about raters such as rating experience in NCLP and activity (i.e., participation in the rating sessions of intermediate-level speaking performances of NCLP), age, gender, and L1. This information was combined with rater severity and consistency indices produced by the Many-Facets Rasch model (MFRM) analysis calculated from operational rating data to create a useful data set for further analysis.

Raters and measures of rater experience

All 104 raters had a university degree in Finnish and the majority were teachers of Finnish as an L1 or L2. They had participated in obligatory rater training approved by the National Agency for Education, which is a prerequisite for becoming an official rater in the NCLP test system. Almost all raters had Finnish as an L1 (with the exception of less than ten raters, whose L1s were German, Estonian, Swedish or Russian), and 94% of

Table 1. Descriptive statistics for different types of rating experience.

	M	*SD*	Minimum	Q1	Median	Q3	Maximum
Years of NCLP rating experience	10.85	7.50	0.01	4.03	9.60	17.66	24.79
Examination sessions rating experience	9.50	6.96	1.00	4.00	8.00	14.00	35.00
Rating volume per session	52.08	15.98	4.89	43.69	50.56	56.31	147.78
Cumulative rating volume	479.88	386.86	4.89	182.50	376.39	684.50	2296.22

them were females. Over the 10 year period, the mean age of the raters was 51 years (*SD*=10; range: 27-70). Since the raters formed a homogeneous group in terms of their education, gender, and L1, it was not meaningful to examine the relationship between these factors and severity and consistency of their ratings.

Raters start their work as an NCLP rater at different phases of their professional careers: some have many years of teaching experience and some only a few. In addition, the raters varied in terms of their rating experience and participation in rating sessions (see Table 1). In this study, the NCLP raters' rating experience is operationalized in four different ways:

(a) "Years of rating experience in NCLP" indicates the number of years for which an individual has been registered with the NCLP as a rater, between 1994 (when the NCLP was first established) and 2019. Therefore, "Years of rating experience in NCLP" increases over time regardless of whether the rater participates in rating sessions or not and can – in theory – range between 0 and 25. There are raters with almost 25 years' experience who have been involved in NCLP from the beginning but other raters who only joined the NCLP during the last year of our research period. The mean years of NCLP rating experience is 11 (*SD* = 8).

(b) "Examination sessions rating experience" indicates the number of different rating sessions (of intermediate-level speaking assessments) for which an individual has participated as an NCLP rater for over the 10-year research period. It can range between 1 and a theoretical maximum of 45, in the sense that there were 45 distinct rating sessions in the period under study (usually, there are 4 to 6 rating sessions per calendar year). Raters usually attend rating sessions at least twice a year. On average, raters participated in 10 distinct rating sessions (*SD* = 7; range: 1–35). One third of the raters attended 9 to 16 times, another third 17 to 35 times, and the remaining third 3 to 8 times.

(c) "Cumulative rating volume" indicates the total count of examinee performances rated by each rater in previous sessions, including the current one, that is, it increases with every rating session the rater attends. The average number of

performances rated by each rater over ten years was 480 (SD = 387; range: 5- 2,296).

(d) "Rating volume per session" indicates the count of examinee performances which were rated by each rater at each rating session. Theoretically, this measure can take values between 1 and the total number of examinees rated under one rating session (which depends on turnout), but in practice the average number of performances rated by each rater per exam was 52 (median = 51) with a standard deviation of 16 and a wide range between 5 and 148.

It is also worth mentioning that when raters participate in rating sessions for NCLP examinations (basic, intermediate and advanced level examinations are administered and rated simultaneously), they do not always rate intermediate-level speaking performances, but they may rate basic- or advanced-level tests or even a different skill, such as writing. In the current study we used data from the rating sessions of intermediate-level speaking performances only.

Measures of disconnect from the CoP

We investigated the effect of the number of rating sessions missed by a rater (referred to as "rater absence") by creating two different variables, borrowing the idea of these variables from Lamprianou et al. (2021). The first was a binary variable ("returning rater"), where code 1 denotes that the rater had attended the previous rating session and code zero denotes that the rater skipped (at least) the previous session. We also counted the total number of rating sessions each of the raters skipped before returning to rate, and we called this variable the "rating gap." Then, we broke down the gap into six ordered categories (referred to as gap_recoded variable) 0, 1, 2, 3, 4, and 5. Each of the ordered categories indicates the corresponding number of skipped rating sessions except for category 5, which means 5 or more skipped rating sessions. The NCLP requires that each rater must participate in rating sessions at least once a year to maintain their license to rate. This means in practice that raters can skip three consecutive tests and must attend the fourth one to keep their license; starting from 2017, as the number of tests increased to six per year, raters can skip up to five tests without losing their rating license.

Many raters attended almost every or at least every second rating sessions but some attended barely one per year. The raters assessed the speaking performances of consecutive tests in approximately 51.7% of cases (i.e., gap_recoded=0). In 22.6% of cases, the raters missed one test (i.e., gap_recoded=1). Raters missed two tests in 10.0% of cases, (i.e., gap_recoded=2), three tests in 4.9% of cases, four tests in 2.8% of cases, and five or more in 8.0% of cases (range: 5-32).

Statistical methods

Our data analysis consisted of two steps. First, we analyzed all our data as a single dataset with the Many-Facets Rasch model (Linacre, 1989) using Facets software (Linacre, 2020) and a three-facets rating scale model (i.e., rater, examinee, and task). To get a unique and comparable severity and consistency measure for each rater at each rating session, we followed the standard procedure described by Lamprianou et al. (2021), Lim

(2009), and Myford and Wolfe (2009). Therefore, to ensure our data connectivity, we linked our data through common examinees (representing over 4% of the total volume of our data), whose language skills did not seem to change across examinations (see more about data linking in Neittaanmäki & Lamprianou, 2024).

We used rater severity measured in logits (the mean rater severity was set to zero; the lower the logit measure, the more severe the rater). Rater consistency was derived from raters' infit and outfit mean squares, produced by the MRMF analysis. Infit mean square (MNSQ) is sensitive to unexpected ratings where the locations of elements are close together on the measurement scale and, respectively, outfit MNSQ where the locations are far apart from each other. Infit and outfit MNSQ values were operationalized as binary variables "misfit," where a value of 1 indicates inconsistency (infit or outfit MSNQ $\geqslant 1.2$) and a value of 0 indicates no inconsistency (infit or outfit MNSQ < 1.2). The results of the MFRM analysis are presented in more detail in Neittaanmäki and Lamprianou (2024). Some MFRM analysis output is displayed graphically in Supplementary Appendix A. Overall, rater severity had small standard errors (0.07 to 0.14) and varied from -2.47 to 2.58 logits. The reliability for rater measures was 0.98. Rxaminee ability ranged from -8.64 for the less able to 9.17 for the more able (examinee ability -1.05 logits; $SD = 3.12$; $SEM = 0.88$). The reliability for examinee measures was 0.90.

Using outfit MNSQ $\geqslant 1.2$ as a cutoff threshold resulted in approximately 14% of raters being classified as misfitting (inconsistent); and infit MNSQ $\geqslant 1.2$ suggested that approximately 11% of raters were inconsistent. Other studies with similar methodologies also used similar cutoff values (e.g., Lamprianou et al., 2021).

To address our research questions, we fit generalized linear mixed models (GLMM) (Agresti, 2013; Bates et al., 2015) using the lme4 package (Bates et al., 2015) of the R platform (R Core Team, 2020). To investigate the effect of different types of experience on rater severity, we fitted four separate GLMMs with rater severity as the dependent variable, but each model had one different type of rater experience as an independent variable. To investigate the effects of different types of experience on rater consistency, we fitted four different GLMMs with rater consistency as the dependent variable, but each model had one different type of rater experience as an independent variable. Typically, all independent variables should be included simultaneously in the same GLMM to avoid multiple testing and account for potential covariates, confounding factors, and interactions. However, in the case of our study, including all types of rater experience in a single model is not feasible, as some of them are highly intercorrelated, and this would lead to collinearity issues (e.g., Pearson's r between "Cumulative rating volume" and "Examination sessions rating experience" variables > 0.9).

To investigate the effect of missing rating opportunities on rater severity and consistency, we used rater misfit as a binary-dependent variable and two different measures of missing rating opportunities as independent variables.

Results

The effect of different types of experience on severity and consistency

Firstly, we used a GLMM with rater severity as a continuous dependent variable and each of the different types of rating experience as independent variables: years of rating experience in NCLP, examination sessions rating experience, rating volume per session, and cumulative rating volume. The raters were modeled as random effects.

Table 2 presents the results of the models explaining rater severity by years of NCLP rating experience, examination sessions rating experience, and cumulative rating volume. As shown in Table 2, the coefficients of the three variables were statistically significant. The coefficient of years of NLCP rating experience (i.e., simply being registered with the NCLP) was statistically significant and positive (0.03, $p < .001$), albeit small. This suggests that raters who were registered with the NCLP for a longer period were more likely to be more lenient. An alternative but equally plausible explanation is that raters become more lenient just because time passes and they remain affiliated with the NCLP as (active) licensed raters. For each year, they remain affiliated licensed raters with the NCLP, and they become more lenient by 0.03 logits. For a period of 25 years (the maximum theoretical possible value), that would be the equivalent of 0.75 logits or around 24% of the standard deviation of candidate ability. It is important to note that the impact on leniency is very gradual, in the sense that one additional year of experience with the NCLP only increases leniency by a small amount (i.e., 0.03 logits). In other words, the increase is so smooth that it is likely difficult for the casual observer (i.e., individual raters) to detect in real time and only becomes practically meaningful and noticeable after longer periods.

On the other hand, the coefficient of examination sessions rating experience was also statistically significant and positive (0.02, $p < .001$), albeit also near zero. This suggests that accumulating experience by participating in rating sessions (not by merely being registered with NCLP) made the raters slightly more lenient over time. However, the magnitude of the effect was small. For example, after participating in 45 rating cycles (the maximum theoretical possible value), the increase in rater leniency is about 29% of a standard deviation of the examinee ability. This, in combination with the findings described in the previous paragraph, highlights the value of longitudinal datasets that allow researchers to quantify smaller changes over time, which would otherwise go unnoticed.

Rating volume per session, the number of performances rated by each rater per each examination did not affect rater severity ($p = .695$; so this is not shown in Table 2), but the coefficient of cumulative rating volume was statistically significant (0.11, $p < .001$), suggesting that the raters became more lenient when they rated more performances. The value of the coefficient in Table 2 refers to the standardized cumulative rating volume ($M = 0$ and $SD = 1$), suggesting that a rater who rated one standard deviation more performances (compared to the mean) would be more lenient by 0.11 of the standard deviation of candidate ability. Again, this indicates only a gradual change in leniency, suggesting that practically impactful changes in severity are less likely to be quantified, unless a long-term longitudinal design is used.

Table 2. The results of the models predicting rater severity.

Predictors	Severity measure			Severity measure			Severity measure		
	Estimates	95% CI	p	Estimates	95% CI	p	Estimates	95% CI	p
(Intercept)	–0.24	–0.37 – –0.11	<**0.001**	–0.11	–0.22 – –0.01	**0.032**	0.03	–0.06 – 0.13	0.464
Years of rating experience in NCLP	0.03	0.02 – 0.04	<**0.001**						
Examination sessions rating experience				0.02	0.01 – 0.02	<**0.001**			
Cumulative rating volume (STD)							0.11	0.06 – 0.15	<**0.001**
Random Effects									
σ^2	0.45			0.45			0.45		
$\tau_{00\ rater}$	0.18			0.19			0.19		
ICC	0.28			0.29			0.30		
$N_{\ rater}$	104			104			104		
Observations	1,418			1,418			1,418		
Marginal R^2/ Conditional R^2	0.056/0.323			0.019/0.305			0.017/0.308		

Note: Positive coefficients suggest an increase in leniency (or a decrease in severity).

Table 3. The results of the model predicting rater (in)consistency by rating volume per session.

Predictors	Outfit12			Infit12		
	Odds Ratios	95% CI	p	Odds Ratios	95% CI	p
(Intercept)	0.09	0.06–0.14	**<0.001**	0.06	0.03–0.09	**<0.001**
Rating volume per session (STD)	0.65	0.51–0.84	**0.001**	0.65	0.49–0.86	**0.003**
Random Effects						
σ^2	3.29			3.29		
$\tau_{00\ rater}$	1.94			2.65		
ICC	0.37			0.45		
$N_{\ rater}$	104			104		
Observations	1,418			1,418		
Marginal R^2/ Conditional R^2	0.033/0.392			0.030/0.463		

Cumulative rating volume and examination sessions rating experience are entwined in the sense that the more often the rater attends rating sessions, the more performances are rated by the rater, so they may not be included in the same model as covariates (Pearson's *r* above 0.9). When the variables of "examination sessions rating experience" and "years of rating experience in NCLP" are included in the same model, the coefficients of both variables remain statistically significant and consistent with the previous models discussed above; the Pearson correlation between the two variables is $r(1416) = 0.407$, $p < .001$ (see Supplementary Appendix B for details).

Intraclass correlation (ICC), the ratio of the between rater variance to the total variance, can also be interpreted as the correlation among observations within the rater. In our models, ICCs were moderate (0.28–0.30), suggesting that the severity varied moderately within the rater (i.e., there is no clear consistency among a rater's leniency/severity).

Rater consistency was modeled as a binary-dependent variable in different GLMMs (1 = inconsistent and 0 = consistent rater). Years of rating experience in NCLP ($p = .711$ for outfit and $p = .745$ for infit MNSQ), examination sessions rating experience (rating frequency) ($p = .960$ for outfit and $p = .472$ for infit MNSQ), and cumulative rating volume ($p = .717$ for outfit and $p = .705$ for infit MNSQ) did not affect rater consistency (so their results are not shown in Table 3 for the sake of brevity). However, as shown in Table 3, statistically significant ($p = .001$ and .003) odds ratio values of 0.65 were found for the independent variable of rating volume per session. This suggests that raters who rated more performances per rating session were less likely to be classified as misfitting/inconsistent.

The effect of disconnecting from the CoP

GLMMs were used to investigate how absences (ordered categorical variable) affected severity and consistency. As before, the raters were modeled as random effects.

Table 4. The results of the model predicting rater (in)consistency by the number of missed rating sessions.

Predictors	Outfit12			Infit12		
	Odds Ratios	95% CI	p	Odds Ratios	95% CI	p
(Intercept)	0.08	0.05–0.12	<**0.001**	0.05	0.03–0.09	<**0.001**
gap_rec [missing 1]	1.07	0.67–1.71	0.768	1.25	0.75–2.09	0.392
gap_rec [missing 2]	1.83	1.03–3.24	**0.038**	1.19	0.60–2.36	0.616
gap_rec [missing 3]	0.77	0.29–2.02	0.590	0.91	0.32–2.64	0.865
gap_rec [missing 4]	3.05	1.15–8.08	**0.025**	2.88	0.97–8.56	0.057
gap_rec [missing 5 or more]	2.04	1.07–3.90	**0.031**	1.69	0.80–3.59	0.169
Random Effects						
σ^2	3.29			3.29		
$\tau_{00\ rater}$	1.88			2.41		
ICC	0.36			0.42		
$N_{\ rater}$	104			104		
Observations	1,314			1,314		
Marginal R^2/ Conditional R^2	0.018/0.375			0.009/0.428		

Missing/skipping rating sessions ($p = .301–.830$) did not seem to explain rater severity (results not shown in Table 4 for brevity). However, missing/skipping rating sessions seemed to affect rater inconsistency significantly (see Table 4). Three out of five odds ratio values for the outfit MNSQ measures, were statistically significant and greater than one, indicating that if a rater missed more than one rating session, they were more likely to be classified as misfitting/inconsistent. There was no statistically significant effect for the infit MNSQ measures.

In conclusion, the results of the effects of rater absence on rater consistency are small but not negligible. Contradictory results of the effect of rater absence on outfit and infit MNSQ can be explained by the actual nature of the two measures. Rater inconsistency measured by outfit MNSQ can, in practice, mean that an otherwise severe rater may give an exceptionally good mark to a poorer performer, or an otherwise lenient rater may give an exceptionally poor mark to a good performer. On the other hand, the infit MNSQ measures are less affected by such extreme rating behavior.

Discussion

In this article, we have examined the effects of rating experience on rater severity and rater (in)consistency, in a high-stakes, Finnish as an L2 speaking examination. In addition, we investigated how gradual disconnection from the CoP may affect rater severity and consistency.

We have broken new ground in speaking assessments by analyzing a granular, longitudinal rater dataset, covering 45 rating cycles. Briefly, we found that with experience comes leniency. However, different measures of experience have quantitatively different magnitudes of impact or no impact at all. We have also demonstrated that raters who rate more performances per examination tend to be less likely to be classified as misfitting; in other words, raters who are more hesitant to rate more performances are also more likely to be classified as inconsistent. Finally, and this is possibly the most interesting of our findings, we have shown that skipping rating opportunities increases the likelihood of a rater to be classified as significantly inconsistent. Although we expected to see a more gradual alienation from the CoP, it seems that skipping even a few rating opportunities (e.g., just two) increases the likelihood for a rater to be classified as significantly inconsistent.

In more detail, our results suggest that while raters become more experienced, they also tend to become more lenient. Although the effect was moderate rather than large, our findings accord with those from previous research. For instance, in Ahola (2016, 2022), raters reported that through experience, they learned to better tolerate inaccurate Finnish and to understand learners from different backgrounds better. Moreover, according to their own perceptions, they learned to view language skills more holistically than they did at the beginning of their rating career. Other studies—but with writing assessment—have also shown that experienced raters tend to assess more holistically (see e.g., Barkaoui, 2010b, 2011 and Ahmadi Shirazi, 2019). Weigle (1999) and Song and Caruso (1996) also found that more experienced raters may be more lenient.

Comprehensibility of oral speech has become increasingly important as the ethnic diversity of immigrants has grown in Finland, and, thus familiarity with different pronunciations may be important. This may be one reason for the raters in our findings becoming more lenient as they gain more experience in rating examinees of a particular profile– that is, when they become more familiar with different ethnic groups speaking in Finnish. Many previous studies (e.g., Carey et al., 2011; Winke et al., 2012) have stated that this kind of rating experience (i.e., familiarity with particular examinees) makes scoring easier, and it can also lead to leniency in scoring. If rater experience with examinees of a particular profile is an important determinant of severity/leniency, then it makes sense for policymakers to invest in training and also retaining pools of raters with homogeneous rating experiences. Sudden changes in the pool of raters by recruiting too many new raters with different rating experiences may lead to what Lamprianou et al. (2021) have described as "cultural shocks" within the CoP, undermining the quality of rating. On the other hand, the risk of too much familiarity with examinees' accent is that raters may understand the performance of an examinee too well compared to ordinary people, and, thus, this can lead to situations in everyday life where the examinee fails to cope with their language skills as expected. This may be an example of rater experience having an unintended and probably negative—rather than positive—consequence.

It is also reasonable to suggest that changes in the examinee group would affect the raters and the general line of rating, in our case, toward leniency. The changes in the group of NCLP examinees and the relatively large increase in the number of examinees over time reflect global events and subsequent processes, such as migration. Forced migration background may lead, for example, to attending the test prematurely in

relation to one's proficiency because of an immense need to apply for citizenship. In the case of the NCLP, it is normal for candidates to retake the test multiple times until they succeed: more specifically, approximately one-third of the examinees retook the test at least once during the last nine examination rounds of the research period. This phenomenon is not unique in the context of Finland; it has also been observed in other cases, such as Australia (Hamid et al., 2019) where language tests have allegedly been used as "gatekeeping tools" (p. 226). The phenomenon is not new either as there are pertinent references on retakes of language tests in the United States as early as 1998 (Del Valle, 2003). The increase in examinees with significantly lower general ability, combined with the fact that the raters have become more familiar with examinees from a wider range of different backgrounds, may have influenced the raters to become more lenient in general. Given that the pool of candidates may be of lower overall ability, it is possible that raters have gradually shifted their attention to different features of performance (Pollitt & Murray, 1996).

However, the same factors did not seem to affect rater inconsistency. What turned out to influence rater inconsistency was absence from rating sessions, that is, how many rating sessions raters skipped before returning to rate again. This finding is consistent with previous research suggesting that disengaging from the CoP increases the likelihood of a rater producing unreliable ratings (Lamprianou et al., 2023). In our context, missing one rating session did not seem to have a significant effect on raters' odds of yielding a misfitting rating pattern, but missing more than one seemed to considerably increase the probability of being classified as inconsistent. The gradual effect of rater absences on consistency may be explained by the fact that the raters are indeed working with similar learners (as test-takers) in their everyday lives, and, therefore, raters are not really alienated from the test-takers and ratings immediately; alienation is gradual. Thus, after an extended break in conducting ratings, uncertainty about the interpretations of the tasks and rating criteria as well as unusual performances can have a stronger influence on the consistency of ratings, especially at the beginning of the rating session. This interpretation may be supported by the fact that the results also suggested that the rater was less likely to be classified as misfitting/inconsistent if they did more ratings during one rating session. If the number of ratings is large enough, it may be easier for the raters to find and stick to their "own rating line" and, thus, feel more confident in their ratings. Lamprianou et al. (2023) also suggested that more confident raters were less likely to be flagged as inconsistent.

Inconsistency in ratings can also be due to the fact that raters cannot maintain a uniform level of severity across the performances that the rater is assigned to assess. In other words, raters may assess some examinees with certain background factors more harshly or leniently. For example, this may happen because of unfamiliarity with a certain group of examinees (e.g., different ethnicities/pronunciations; Kang et. al., 2023). Often, the rater must first get used to the foreign accent to understand better what the speaker is communicating. In addition, based on our experience in the operational test system, we have found that quick and good decision-making skills may be linked to consistent rating behavior. A rater who must think long and hard about their rating and listen to the performance several times is more likely to prove inconsistent. This suggests that intra-rater reliability is probably related to decision-making skills (which can be improved by

training, up to a point). This should be taken into account when selecting raters for the test system but also when designing rater training sessions.

Finally, the test system should ensure that raters assess regularly without excessive breaks and that the number of assessments is large enough and has sufficient variation (i.e., performances from different L1 speakers). Mandatory rater training (where raters can discuss with colleagues) before every rating session is important because that way, raters recall the rating criteria and listen to the benchmark samples regularly and finally are not alienated from the CoP (see Lamprianou et al., 2023). Only through regular training can the test system improve the consistency of the ratings and at the same time reduce the variation in severity between raters. Otherwise, it is difficult for the test system to address the leniency that comes with experience and the inconsistency that comes with alienation from the rater community.

Conclusion

In this article, we investigated the effect of different types of rater experience on raters' severity and consistency in the context of a high-stakes intermediate-level speaking test in the NCLP examination system in Finland. We found that with experience comes leniency, but different experience measures have quantitatively different magnitudes of impact. We showed that skipping rating opportunities increased the likelihood of a rater to be classified as inconsistent. We also demonstrated that raters who rate more performances per examination are less likely to be classified as misfitting. Based on our findings, we recommend that the test systems should ensure that raters assess regularly without excessive breaks. Understanding the effect of alienation from the rater community is important for ensuring the validity and fairness of the ratings, particularly in high-stakes contexts.

Though we have shown here that both training and rating experiences within the test system are relevant aspects in the rating, raters' teaching experience (e.g., Royal-Dawson & Baird, 2009), educational and professional background (e.g., Shohamy et al., 1992), and expectations and attitudes toward examinees with different background factors (e.g., Ahola, 2020; Johnson, 2005) most probably have some kind of an effect on rating behavior. In this study, these factors were not considered. Furthermore, we did not use the experience gained by the rater when rating subskills or test levels other than intermediate speaking as a factor in this study. It will, therefore, be interesting to further investigate how experience gained in different skills and levels is transferred to the assessment of intermediate speaking.

It is reasonable to anticipate that various types of experience will exert varying degrees of influence on rater characteristics. For instance, practitioners need to identify the most influential types of experience to effectively monitor and include them in their rater selection criteria. One approach to assessing the distinct impact of independent variables on the dependent variable was to sequentially add all relevant variables to the same model. Regrettably, this was only partially feasible with our dataset (see Supplementary Appendix B), due to high correlations among some independent variables. In the future, it is important to construct appropriate datasets to facilitate this type of research. There also seems to be a knoweldge gap regarding the micromechanisms by

which some variables (e.g., different types of experience) may or may not affect the consistency and severity of the rating. Unfortunately, such research questions are impossible to answer using quantitative data alone. In the future, we suggest researchers approach these types of gaps utilizing more in-depth studies (e.g., more qualitative data) with rich techniques such as focus groups, interviews, or think-aloud protocols.

Finally, we have a methodological recommendation for future research. Our study has benefitted significantly from a longitudinal dataset, one spanning 45 consecutive testing cycles, to detect small effects on consistency and severity. The coefficients of the models were small, which suggests that datasets covering small numbers of consecutive testing cycles would likely fail to detect measurable signals. From the perspective of researchers, it is imperative to form partnerships with testing organizations which are in possession of longitudinal data of this size. Similar thoughts about the usefulness of longitudinal designs have recently been expressed by other researchers as well (Lamprianou et al., 2021). We encourage the research community to invest more resources in longitudinal designs in the future to be able to investigate more subtle and gradual changes in rater effects.

Acknowledgements

We 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. We 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: This research was conducted as part of the Broken Finnish project funded by the Research Council of Finland (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

Supplemental material

Supplemental material for this article is available online.

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.

Ahmadi Shirazi, M. (2019). For a greater good: Bias analysis in writing assessment. *SAGE Open*, *9*(1), 2158244018822377. https://doi.org/10.1177/2158244018822377

Ahola, S. (2016). Puhetta arvioinnista: Yleisten kielitutkintojen arvioijien käsityksiä arvioinnista [Raters' beliefs and views of assessment in the National Certificates of Language Proficiency]. In Huhta, A. & Hildén, R (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. (2020). Sujuvaa mutta viron kielen vaikutusta: Yleisten kielitutkintojen arvioijien käsityksiä vironkielisten suomenoppijoiden suullisesta taidosta [Fluent but influenced by Estonian: Rater perceptions of the spoken Finnish skills of L1 Estonian speakers in National Certificate exams]. *Virittäjä*, *124*(2), 217–242. https://doi.org/10.23982/vir.79831

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

Alp, P., Epner, A., & Pajupuu, H. (2017). The influence of rater empathy, age and experience on writing performance assessment. *Linguistics beyond and within*, *3*, 7–19. https://doi:10.31743/lingbaw.5647

Ang-Aw, H. T., & Chuen Meng Goh, C. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal*, *42*(1), 31–51. https://doi.org/10.1177/0033688210390226

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

Barkaoui, K. (2010a). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44*(1), 31–57. https://doi.org/10.5054/tq.2010.214047

Barkaoui, K. (2010b). 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

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279–293. https://doi.org/10.1080/0969594X.2010.526585

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.org/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

Brennan, E., & Brennan, J. S. (1981). Measurements of accent and attitude toward Mexican-American speech. *Journal of Psycholinguistic Research*, *10*(5), 487–501. https://doi.org/10.1007/BF01076735

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201–219. https://doi.org/10.1177/0265532210393704

Carey, M. D., & Szocs, S. (2023). Revisiting raters' accent familiarity in speaking tests: Evidence that presentation mode interacts with accent familiarity to variably affect comprehensibility ratings. *Language Testing*. https://doi.org/10.1177/02655322231200808

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

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

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. https://doi.org/10.1177/026553229000700104

Davis, L. E. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* [Unpublished doctoral dissertation]. University of Hawaii at Mānoa.

Davis, L. E. (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

Del Valle, S. (2003). *Language rights and the law in the United States: Finding our voices*. Multilingual Matters. https://doi.org/10.21832/9781853596445

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and Criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Peter Lang.

Erdosy, M. U. (2004). Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions. *ETS Research Report Series*, *2003*(1). https://doi.org/10.1002/j.2333-8504.2003.tb01909.x

Fan, J., & Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer? *Papers in Language Testing and Assessment*, *8*(2), 117–142. https://doi.org/10.58379/JRWG5233

Hamid, M. O., Hoang, N. T., & Kirkpatrick, A. (2019). Language tests, linguistic gatekeeping and global mobility. *Current Issues in Language Planning*, *20*(3), 226–244. https://doi.org/10.1080/14664208.2018.1495371

Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly*, *13*(1), 25–41. https://doi.org/10.1080/15434303.2015.1134540

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. https://doi.org/10.1080/15434303.2013.769545

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Blackwell. https://doi.org/10.1002/9780470757024.ch15

Khabbazbashi, N., & Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing*, *37*(3), 333–360. https://doi.org/10.1177/0265532219898635

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment* [Doctoral dissertation, Teachers College, Columbia University]. ProQuest Dissertations.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly*, *12*(3), 239–261. https://doi.org/10.1080/15434303.2015.1049353

Knoch, U., Zhang, B. Y., Elder, C., Flynn, E., Huisman, A., Woodward-Kron, R., Manias, E., & McNamara, T. (2020). 'I will go to my grave fighting for grammar': Exploring the ability of language-trained raters to implement a professionally-relevant rating scale for writing. *Assessing Writing*, *46*, 100488. https://doi.org/10.1016/j.asw.2020.100488

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.

Lindemann, S. (2005). Who speaks "broken English"? US undergraduates' perceptions of nonnative English. *International Journal of Applied Linguistics*, *15*(2), 187–212. https://doi.org/10.1111/j.1473-4192.2005.00087.x

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

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

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

Miao, Y. (2023). The relationship among accent familiarity, shared L1, and comprehensibility: A path analysis perspective. *Language Testing*, *40*(3), 723–747. https://doi.org/10.1177/02655322231156105

Mohd Noh, M. F., & Mohd Matore, M. E. E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology*, *13*, 941084. https://doi.org/10.3389/fpsyg.2022.941084

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

Neittaanmäki, R., & Lamprianou, I. (2024). Communal factors in rater severity and consistency over time in high-stakes oral assessment. *Language Testing*, *41*(3).

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System*, *30*(2), 143–154. https://doi.org/10.1016/S0346-251X(02)00002-7

Park, M. S. (2020). Rater effects on L2 oral assessment: Focusing on accent familiarity of L2 teachers. *Language Assessment Quarterly*, *17*(3), 231–243. https://doi.org/10.1080/15434303.2020.1731752

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 74–91). University of Cambridge Local Examinations Syndicate and Cambridge University Press.

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

Royal-Dawson, L., & Baird, J.-A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice*, *28*(2), 2–8. https://doi.org/10.1111/j.1745-3992.2009.00142.x

Şahan, Ö., & Razi, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing*, *37*(3), 311–332. https://doi.org/10.1177/0265532219900228

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

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing*, *5*(2), 163–182. https://doi.org/10.1016/S1060-3743(96)90023-5

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. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6

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

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150–173. https://doi.org/10.1016/j.asw.2011.12.001

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35–51. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=e7abfd4eae82a75fd542054f078485bd924f8da1

Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment*, *2*(1), 37–56. https://escholarship.org/uc/item/83b618ww

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x