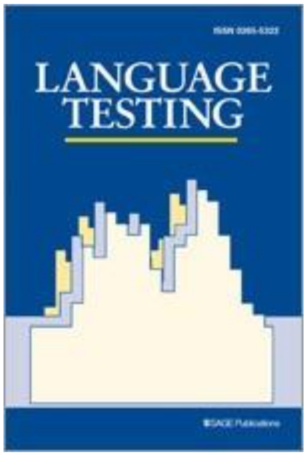


The stability of rating characteristics in high-stakes English as a Foreign Language examinations: methodological and substantive considerations

Journal:	*Language Testing*
Manuscript ID	LT-19-0048.R2
Manuscript Type:	Original Manuscript
Keywords:	raters, rating characteristics, stability of rating, severity, consistency, rater experience
Abstract:	This longitudinal study (i.e. 2002-2014) investigates the stability of rating characteristics of a large group of raters over time in the context of the writing paper of a national high-stakes examination. The study uses one measure of rater severity and two measures of rater consistency. The results suggest that the rating characteristics of individual raters are not stable, thus predictions from one administration to the next are difficult, although not impossible. In fact, as the membership of the group of raters changes from year to year, past data on rating characteristics become less useful. When the membership of the group of raters is retained, the community of raters develops more stable characteristics. However, 'cultural shocks' (low retention of raters and large numbers of newcomers) destabilize the rating characteristics of the community and predictions become more difficult. We propose practical measures to increase the stability of rating across time and offer methodological suggestions for more efficient rater effect-related research designs and analyses.

DOI: **https://doi.org/10.1177/0265532220940960**

Introduction

High-stakes language exams are used for a variety of purposes, such as the assessment of L1 proficiency (Kuiken & Vedder, 2014), or language proficiency in L2 (Plakans & Gebril, 2018; Winke & Lim, 2015). They typically aim at providing employment opportunities, access to programs, promotion, or even residency permit (Douglas, 2015; Knoch, 2011). Often, they are performance-based exams and rely heavily on expert raters to evaluate examinee responses. Therefore, maintaining raters' stability is an extremely important endeavor that can support the validity claims made by the providers of such exams (Isbell, 2017; Knoch & Chapelle, 2018). Validation of the rating processes, particularly within an argument-based framework, allows also the articulation of a series of inferences that clearly demonstrate how examinees' performance can be interpreted and appropriately used (Kane, 2013; Knoch & Chapelle, 2018).

However, exam bodies are frequently faced with various manifestations of undesirable rating behavior, often referred to with the collective term "rater effects" or "rater errors". Various rater effects have been identified in the literature, but, in this study, we focus only on two: rater severity and rater inconsistency. Severity characterizes a rating behaviour where a rater tends to award higher or lower scores compared to other raters who rate the same responses (Bonk & Ockey, 2003; Kondo-Brown, 2002; Ockey, 2009). Inconsistency refers to raters who demonstrate atypical rating patterns in that their ratings do not align with the rest of the raters because they apply scoring standards inconsistently across responses (Isbell, 2017). A misalignment with the rest of the raters is often fuelled by personal preferences or biases (Wang & Engelhard, 2019), but also by haphazard application of the scoring rubrics among others (see Myford & Wolfe, 2003) which are difficult to predict or account for.

Overall, even though there has been an unprecedented amount of resources invested in rater effects research, some of the pertinent issues remain unsolved. As Author1 (2018) also stressed "[t]he contradicting results extend the confusion and the agony of policy makers and practitioners alike who often receive much pressure by the media and the public […] to have an adequate pool of competent raters in their disposal" (p. 431). For example, instability of rater severity and consistency over a rating period remains a substantive problem. Studies have shown that the rating characteristics change significantly over time (Fitzpatrick, Erickan, Yen & Ferrara, 1998; Hoskens & Wilson, 2001) and even between rating sessions within a few months (Author1, 2006, 2018; Congdon & McQueen, 2000; Harik et al., 2009). However, other studies came up with contradictory results. For example, Lim (2011), Leckie and Baird (2011) and Davis (2016) showed that raters, on some occasions, managed to maintain acceptable rating quality over time.

Unfortunately, research on the consistency of rating characteristics over time is still sparse. Therefore, the aim of this study is to investigate the (in)stability of two rating characteristics severity and consistency) over time, in the context of a writing paper of a high stakes language exam in a European country. The researchers also investigate how the stability of rating characteristics relates to rating experience. The paper concludes by proposing practical measures to increase stability over time and offers methodological suggestions for more efficient research designs in the relevant field.

Literature Review

The topic of rater effects in the area of second language writing assessment has been quite prevalent in the research literature (Slomp & East, 2019). Nevertheless, research has generated mixed results when investigating differences between novice and experienced raters. For example, Weigle (1998) reported that more experienced raters might be more lenient than their less experienced colleagues. However, Bonk and Ockey (2003) argued for the opposite, whereas Barrett (2001) reported mixed results. Interestingly, Huhta, Alanen, Tarnanen, Martin and Hirvela (2014) claimed that, despite variation in severity due to raters' experience, nearly all of them rated consistently enough to be trusted as raters. However, there are important qualitative (e.g. see Barkaoui, 2010a, 2010b) as well as quantitative differences (Deygers & Van Gorp, 2015) between raters of different experience.

Raters' training also attracted a significant number of studies. Elaborated training sessions were implemented by Elder, Barkhuizen, Knoch and von Randow (2007), who claimed that training may have had only a minimal impact in terms of severity and consistency; instead, training made raters overcautious. The limited effect of training on severity and consistency was also reported by Davis (2016) and Weigle (1998). Even less optimistic were Han (2015), Knoch (2011) and Lumley and McNamara (1995) who found that ratings were no better in terms of severity and consistency when raters were given training.

Experience and training seem to be important, but the mixed research findings suggest that reality may be more complicated than it seems and that other factors might be at play, e.g. language, educational and professional background of raters (see Lim, 2011). More recently, an interesting tendency has been observed among raters that of gradually developing their own Community of Practice (CoP) across time that emerge when raters immerse in the culture of the group and internalize the rules of engagement (Al-Maamari, 2016; Herbert, Joyce & Hassall, 2014). Wenger, McDermott and Snyder (2002) consider a CoP as a "group of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in the area by interacting on an ongoing basis" (p. 7). The notion of a CoP has recently featured in the assessment literature and promotes that individual raters develop an understanding of the rating criteria and norms as a social constructivist process of learning (Willey & Gardner, 2011). Such learning often happens during official standardization sessions through discussions, collaborative re-rating of scripts to reach consensus, feedback sharing and reflection on rating practices. To understand CoPs, Herbert, Joyce and Hassall (2014) suggested standardization meetings of raters as dynamic and developmental social activities, grounded in the situated practice of the group.

However, learning also happens through any informal interaction (e.g. during breaks or short exchanges in the corridors of marking centers) where raters have the opportunity to (a) learn through the community by belonging to the group; (b) practice, i.e. learn by doing; (c) ascribe meaning to their practice through experience; and (d) develop identity (see also Herbert et al., 2014). For example, in a recent community-building exercise, Willey and Gardner (2011) reported quotes from raters which show how the collective culture of the group shapes the practice of individuals: "I was able to see what they […] were thinking … [and] …learn and improve my own technique … [and] … I was able to get a feel for how others mark ... a learning experience" (p. 669).

Shay (2005) also suggested, drawing on Bourdieu's theory of social practice, that achieving consistency across the group means achieving consensus. Thus, when rater retention is low and many new raters join the group, it is reasonable to expect a change in the culture and the shared understandings and practices and, therefore, lack of consistency among raters. It is possible that higher rater retention can contribute to the creation of a more stable community of raters with a more solid shared understanding of the rating system that can lead to higher rater consistency. On the contrary, radical changes in the membership of the group of raters could cause a cultural “shock” which could destabilize the community while time may be needed to renegotiate the rating standards and their practices through formal training and informal interaction among them.

The issue that needs to be investigated is the degree to which radical changes in the membership of the group of raters may cause major shifts in the modus operandi of the CoP. It is the tenet of this research that the cohesion of the community is not a function of the sum of the experiences of the individual raters, but rather the minimum shared experience of the group. In the context of this particular study, the prevalence of the collective wisdom of the CoP is even enshrined (implicitly) in the country’s Examination Law, which specifies that exam papers need to undergo two blind ratings. In case the two scores differ by more than 10%, then a third rater is called in for another blind rating to resolve the discrepancy. This is a clear example of a system where the collective wisdom (e.g., the average of triple blind ratings) prevails over the concept of the “gold standard” (that is, the judgment of a Principal Examiner or the judgment of a very small group of “senior” raters).

The concept of the ‘gold standard’ has prevailed in many settings, not only in educational assessment (Leckie & Baird, 2011) but also in other disciplines, such as medicine (e.g., Gianinazzi et al., 2015; Haj-Ali & Feil, 2006; Zhang et al., 2012). Contrary to the “gold standard”, in the context of our study, the community of raters defines the standards through elaborate designs of multiple blind ratings and standardization meetings. This model can be found in operation in geographically and culturally diverse countries, such as Australia, Canada, China, Cyprus, Greece and Malta (see MATSEC Support Unit, 2018; NSW Government, 2019).

Context and Aims of the Study

This study investigates the longitudinal stability of two rating characteristics (severity and consistency) in the context of the operational marking of the paper-and-pencil writing paper of a high-stakes exam of English as a Foreign Language (EFL) in a European country. The EFL exam is part of a wider battery of tests taken by thousands of students every year aspiring to progress to higher education. Overall, the university-entrance exam is high-stakes, in the sense that there are more candidates than the number of available places[1].

The study aims to answer the following questions:

RQ1: To what extent is the severity and the consistency of individual raters stable across time?

RQ2: To what extent are radical changes in the composition of the group of raters related to changes in severity and consistency?

RQ3: To what extent is rating experience related to raters' severity and consistency?

RQ4: To what extent is it practically informative to use past data as a proxy of future rater severity and consistency?

Drawing on past research, we hypothesize that our ability to use past data as a proxy for the future rating characteristics of the raters will heavily depend on rater retention (we use "rater retention" to refer to the degree to which the group of raters does not change over time). In the years where retention is high (when the composition of the group of raters is very similar to that of the previous year), we expect the concordance between the rating characteristics of the raters on consecutive years to be higher.

The dataset used in this study has the format of a repeated measures design. Each of the datasets contributes one observation for each of the two rating characteristics (severity and

[1] the data used in the current study was made available on the condition of non-disclosure of the name of the exam organization or the country

consistency) for each of the raters in the dataset. The two rating characteristics were analysed independently, but the questions are the same.

Methods

The exam and the data

The study uses thirteen EFL examination datasets from 2002 to 2014. All datasets come from the same EFL test of the competitive high-stakes university entrance exam.

The EFL examination consists of four components (writing, reading, language usage and listening but not speaking) (see Appendix A). The current study focuses on the writing component of the exam. The writing task typically expects the candidates to generate a text of 250-300 words. The raters use a pre-specified rubric which consists of five criteria: content, organisation, vocabulary, grammar and mechanics. Not all criteria have the same number of descriptors or worth the same marks (see Appendix B). There are not any automatic score reductions if the examinees write a shorter or a longer essay; this is left to the judgment of individual raters. Overall, the maximum possible score on the writing task was 25 points for the years 2002-2011 and 30 points for the years 2012 until 2014.

Finally, it is worth mentioning that the pool of candidates and the content of the examination have not changed substantially over time.

The rating procedure

The essays of the candidates for the writing task are rated independently by two raters. Therefore, the final score of any examinee is the sum of the total scores awarded by the two raters. In cases where the total scores differ by 10% or more, then a third rater is invited to rate the responses independently. The final score for the examinee in such cases is the sum of the three ratings multiplied by 2/3.

Before the operational rating, the raters undergo a one-day formal standardization meeting which includes very detailed guidelines and instructions. The raters study the marking scheme, rate a number of sample scripts for standardisation purposes and discuss ways to avoid being too severe or lenient. During the process, the marking scheme can be modified or improved to satisfy the community of raters, resulting also in a sense of ownership among its members. When the officer in charge of the operational rating decides that the standardization meeting has concluded, the operational rating begins (usually on the next working day). Both the training and the operational rating take place in a centralized location. The raters share the same working offices, are allowed to interact and consult each other and they typically share the same space for coffee breaks. The operational rating usually lasts a whole working week during which the raters typically discuss their rating experience and exchange ideas and comments regarding the performance of the examinees.

Examinees and raters

Most of the examinees were 18-year-olds graduating from upper secondary education. A very small percentage took the exam for a second time (rarely for a third time). During the years 2002-2014, a small proportion of around 7% of the examinees repeated the exam to improve their scores in order to gain access to popular university departments. The number of examinees ranged from 1569 to 3237.

The number of raters ranged from 19 to 28 per dataset (24 raters on average per year). The raters are qualified secondary education EFL teachers. Any teacher working in a public upper secondary school may apply to become a rater, as long as they satisfy specific criteria (e.g. years of school experience). At any given year, the average experience of the raters with this particular examination was around 4.5 years (sd=4.2). However, some raters have rating experience with other, similar, examinations set by the same organization. The average experience of the raters with those examinations is around 1.4 but the distribution is highly skewed (see Appendix C for more information).

Each rater marked, on average, approximately 170 scripts per exam. There were 99 different individual raters in total and each of them participated in a different number of exams (see Table 1). One rater participated in 12 exams, two raters participated in 11 exams and other raters participated in a smaller number of exams. Overall, each rater participated in approximately three exams.

[Insert Table 1]

Although consecutive exam years draw from the same pool of raters, exam years further apart share very few raters, if any. For example, there are 14 common raters between exam years 2012 and 2013, but there is only one common rater between years 2002 and 2013.

As it can be seen in Table 2, consecutive exam years share at least eleven raters. The proportion of raters who rated in two consecutive years ranged from 39% to 84%. On average, 65% of the raters at any given year also rated scripts in the preceding year.

[Insert Table 2]

The raters, however, do not always participate in blocks of consecutive years. They sometimes have a break of one or more years, and then they come back to rate scripts in subsequent years. Such an instance is rater 1, who participated only in years 2006 and 2008 (skipped 2007), whereas rater 2 participated in the years 2004 and 2005, had a break of two years and then rated again in 2008. For any given examination, we will hence use the term 'returning rater' to denote the raters who had also rated in the immediately previous year (e.g. for the 2008 dataset, a rater is 'returning' if he/she had also rated in 2007).

Data analysis and key variables

For the identification of rater effects, we used the Many-Facets Rasch model (MFRM) (Linacre, 1994). For the sake of brevity, we will not present the relevant formulas, but more information about MFRM is available in Colleagues and Author1 (2018) and Author1 (2020). For the analysis of the data we used the Facets software (Linacre, 2005).

Data connectivity

A fundamental requirement of MFRM is to identify "common" elements for at least two of the three facets of measurement (e.g. items and examinees) between datasets. This is known as the "connectivity" requirement, that is when the elements must be linked so that there are no 'disjoint subsets' (Lim, 2011; Linacre, 1989).

Within each exam year, our datasets are strongly connected by design. Unfortunately, there is no data connectivity across time because the examinees are asked to write a different essay every year. Also, although there are examinees who repeat the examination in subsequent years, it is not straightforward to assume that they are qualitatively the same person because a whole calendar year elapses between consecutive examinations.

Because of the lack of connectivity, we firstly analysed each year's dataset independently, so we run 13 independent MFRM analyses. We used this analysis to evaluate the model-data fit and to get a general feeling of the data (we will discuss these findings in the next section). However, analysing each dataset separately does not provide a common reference within which to directly compare individual rater severity over time. Thankfully, Lim (2009, 2011) and Myford and Wolfe (2009) provide very meticulous instructions to generate a linked design from disconnected datasets such as ours. Therefore, we will only describe the main steps of the procedure for the convenience of the reader as follows:

1. We firstly identified the individuals who repeated the examination in two consecutive years (e.g. 226 candidates repeated the examination in years 2002 and 2003).
2. Then, we compared the performance of each of those examinees, on the two consecutive years, using components of the examination other than writing which is the focus of this study, (e.g. "reading comprehension", "language use"; see Appendix A). Conveniently, there was a very high correlation between the performances of the candidates for any two consecutive examination years[2]. This supports the argument of the "fossilization of abilities" and suggests that it is not unjustified to treat a common examinee between two years as being practically the same person (e.g. see Lim, 2009).

[2] For example, for the years 2002/2003, for the "listening" component of the examination, the correlation of the scores of the common candidates was 0.55 (p<0.001) for the first 10-point section and 0.63 (p<0.001) for the second 10-point section (a total of 20 points for "Listening"). The corresponding correlations for the years 2003/2004 were 0.45 (p<0.001) and 0.52 (p<0.001); for the years 2004/2005 the correlations were 0.60 (p<0.001) and 0.67 (p<0.001) etc. The corresponding correlations for the two sections of the "Language use" component of the examination were, for years 2002/2003, 0.59 (p<0.001) for the first section (5 points) and 0.68 (p<0.001) for the second section (10 points) for a total of 15 points for "Language use"; for the years 2003/2004 the corresponding correlations were 0.48 (p<0.001) and 0.65 (p<0.001) etc.

3. We considered an examinee to be "common" between two consecutive examination years, only if their performance was the same (within the nearest whole mark).

After the matching procedure was completed, the whole dataset was connected over the thirteen exam years, permitting us to run a single MFRM. For each rater, for each year, the analysis yielded a directly comparable severity estimate (in Rasch logits).

Data analysis

We operationalize "consistency" using the Rasch fit statistics. For each rater, the MFRM yielded two measures of model-data fit: Infit Mean Square (IMS) and Outfit Mean Square (OMS). High values indicate a more misfitting (= inconsistent) rater. The OMS is inflated (compared to the IMS) by unexpected ratings which are off-target (for example, when an otherwise very able examinee is awarded a very low mark on an easy item).

IMS and OMS theoretically take values from 0 to infinite. Values lower than 1 indicate overfit (there is too little stochasticity in the data), thus they are not detrimental to the quality of measurement. Values above 1 indicate misfit (which is an indication of inconsistent or haphazard rating) and are threatening as they imply that the model does not describe the data well. There is a huge body of literature about the cut-off values for IMS and OMS that may indicate too much misfit. We do not believe that there should be catholic cut-off values for these indices to be used in every context. Therefore we decided to use four different cut-offs: 1.15, 1.2, 1.25 and 1.3, in accordance to Author1 & Colleague (2004) and Author1 (2020, p.250). Our goal is to provide more information to the reader and to avoid potential criticisms of unintentionally affecting our findings. The cut-offs of 1.3 and 1.2 are widely used in the literature; we also used the cut-offs of 1.15 and 1.25 to give more information to the readers as model-data fit is a continuum rather than a discrete phenomenon. For each dataset, we created four variables – one for each cut-off value - which were coded as 1 (indicating significant misfit for the individual raters) or 0 (indicating insignificant misfit for individual raters). To improve the readability of the manuscript, we only present tables and figures for the widely used cut-off value of 1.2. Figures and models for all four cut-off values, however, are presented in the Appendices.

The MFRS also provided a measure of the severity of each rater, for each dataset. Rater severities are measured in Rasch logits, in a scale typically spanning from -2 to 2.

Finally, we used three different measures of rating experience: a) the variable "Experience" denotes the years of rating experience for this specific exam program, b) the variable "Other Experience" represents the years of rating experience in other language assessment programs, and (c) the dichotomous variable "Returning rater" signifies rating in the immediately previous exam cycle (1=retuning rater vs 0=non-returning rater). The correlation between the "Experience" and "Returning rater" variables is around 0.4, so collinearity was not a threat.

Changes in severity and consistency were analysed using graphical methods (e.g., scatterplots) as well as Generalized Linear Mixed Models (GLMMs, see Author1, 2020). To investigate the relationship between the rater statistics for consecutive years, we used Pearson correlations (for numerical variables) and chi-square tests (for dichotomous data).

Results

Many-facets Rasch model-data fit

As mentioned above, initially we analysed each of the thirteen datasets independently. Given the confines of space, it will not be possible to discuss the thirteen analyses in depth. However, some summarized information is presented in Appendix D which shows the average values and standard deviations (in parentheses) of the model-data fit for each of the thirteen datasets. The last column shows the separation index (and the reliability in parenthesis; higher values are better on a scale from 0-1) as well as the number of examinees. The model-data fit was satisfactory, for all datasets and for all practical intents and purposes of this study. As expected, there were some candidates' responses misfitting the MFRM but these were not considered capable of actually invalidating the meaningfulness of the Rasch measures. The separation indices for the measures of the raters were very high (see last column of Appendix D). This is a strong indication that there is significant variability between rater severity within each dataset. A few of the items (criteria) also had larger than expected fit statistics but the model-data fit is overall satisfactory.

The thirteen datasets were collated in a connected dataset as described in a previous section. The connected dataset was analysed using the MFRM in a single analysis, having approximately 25,500 different examinees. The separation index (and reliability) is satisfactory, suggesting that the examination can effectively differentiate between the examinees. The mean examinee ability was -0.32 logits and the standard deviation was 2.98. The mean IMS was 0.93 and the mean OMS was 1.03.

The mean severity of the raters was set to zero (SD=0.74) with a range of around 5 logits (from -2.86 to 2.18). The precision of measurement for the raters was very high, with a separation index (e.g. reliability) of 12.86. The average of the standard error of the rater Rasch measures was extremely small, between 0.03 and 0.06, suggesting a high precision of measurement, mainly because of the large sample size of examinees involved in the analysis. The model-data fit for the raters was generally satisfactory; the mean IMS of the raters was 0.97 (SD=0.23) and the mean OMS was 1.10 (SD=0.53).

The mean difficulty of the items (i.e, the criteria) was -1.96 (SD=0.99), with a mean IMS of 1.00 (SD=0.26) and a mean OMS of 1.10 (SD=0.43). Overall, the model-data fit was considered to be satisfactory for all practical intents and purposes of the study.

The (in)stability of rating characteristics

This section investigates the degree to which the rating characteristics of individual raters are stable across exams (RQ1).

The Rasch severity measure for each rater, for each exam, is plotted across time in Figure 1 (for clarity, only raters who participated in six or more exams are shown). The x-axis ranges from 0 (the baseline year = Year 2002) to 12 (the last data set = Year 2014). The y-axis represents the Rasch severity estimate for each rater (approximately ranging from -2 to 2 logits). Some of the raters appear to have very stable severity measures, e.g. raters 56 and 77 have a very small standard deviation of Rasch estimates of 0.342 and 0.310 respectively (within each rater, over time). More specifically, rater's 56 severities range from -0.03 to 1.08 logits and rater's 77 severities range from -0.85 to -0.383 logits. Taking into account that the standard deviation of examinee Rasch estimates is approximately 3 logits, we observe that the severity of the most stable raters fluctuates in a range of around a third of the standard deviation of candidates' ability. Other raters seem to have even more volatile severity measures, e.g. raters 75 and 32 have a large standard deviation of Rasch estimates of 1.09 and 1.04 respectively. Rater 75, for example, demonstrates severity estimates which span a range of around 4.2 logits, the equivalent of almost one and a half standard deviations of exainees' ability.

[Insert Figure 1]

Finally, there is significant variability in the probability of different raters to be classified as misfitting. For example, using a cut-off value of OMS ≥ 1.2, rater 50 who had participated in ten exams was never classified as misfitting (Figure 2). On the other hand, other raters were frequently classified as misfitting; for example, rater 85 who had participated in ten examinations, was classified five times as misfitting. Interestingly, some raters are consistently misfitting over blocks of consecutive years (e.g. rater 10) whereas some raters are misfitting in a random-like manner (e.g. rater 77).

[Insert Figure 2]

Figure 3 presents the same information as Figure 2 but for the Infit Mean Square (IMS ≥ 1.2). We observe that there are raters who tend to be classified as misfitting more often, and raters who tend not to be classified as misfitting. For example, rater 50, who had participated in ten examinations, was never classified as misfitting. Other raters, e.g. raters 22 and 32, were much more likely to be classified as misfitting (both were classified as misfitting on half of the occasions).

[Insert Figure 3]

The main conclusion here is that there is a large variability between raters' probability to be classified as misfitting. It is, therefore, interesting to investigate whether experience or other factors may be identified which affect rater severity or consistency.

Rater retention and inconsistency

To answer RQ2, we used the full dataset of 99 raters across the thirteen years. We applied the four cut-off values discussed in the Methodology section to classify each of the raters as inconsistent or consistent in each of the dataset.

There are significant differences between exam years regarding the proportion of raters classified as misfitting. For example, for the OMS cut-off value of 1.2, the proportion of misfitting raters ranges from 8% (Year 2004) to 40% (Year 2007). For the IMS cut-off value of 1.2, the proportion of misfitting raters ranges from 0% (Years 2003, 2004, 2005) to 23% (Year 2012). To investigate the determinants of rater misfit, we used a GLMM where the dependent variable was dichotomous (misfitting / not misfitting rater) and the independent variables were: (a) the proportion of returning raters for the year, (b) the experience of the individual rater on this particular exam (numeric variable), (c) the experience of the individual rater in other exams (numeric variable), and (d) a dichotomous variable indicating whether each of the raters had rated in the immediately previous year (variable name: “Individual rater returning”). Raters are modeled as random effects.

A different model was run for each of the four cut-off values of the OMS and IMS. The table for the four OMS models is presented in Appendix E, but for the sake of readability, Table 3 presents only the results for the second model (cut-off value of OMS=1.2; left part of the table). The main message of Table 3 is that the coefficient of the interaction between the two independent variables (“Proportion of raters returning per year” and “Individual rater returning”) is statistically significant and has a positive coefficient. The probability of a returning rater to be classified as misfitting decreases as the proportion of returning raters in the group increases. In other words, it is much more likely for returning raters to “blend in” their familiar group of raters rather than to rate in an atypical way. On the other hand, the probability of a new rater to be classified as misfitting increases significantly as the proportion of returning raters increases. In other words, new raters may need some effort and time to adopt to the rating culture of the existing group of raters and they are more likely to rate atypically to be classified as misfitting.

[Insert Table 3]

As shown in Figure 4, the probability for a returning rater to be classified as inconsistent decreases considerably as the proportion of returning raters increases. A returning rater has a much smaller probability to be inconsistent if the community of practice is retained. More specifically, the probability of a rater to be classified as inconsistent is almost halved, from around 27% to 17%, as the proportion of returning raters doubles from 0.4 to 0.8. On the other hand, the probability of non-returning raters to be classified as inconsistent increases dramatically as the proportion of returning raters increases. More specifically, the probability increases from circa 8% to 50% as the proportion of returning raters is doubled, from 0.4 to 0.8. Rating experience, either for this particular exam or for other exams did not contribute significantly and is thus not included in the models.

[Insert Figure 4]

We repeated the same analysis using the four cut-off values for the IMS statistic (see Appendix F). We found the same patterns of results as with the OMS statistic (see Appendix G for a graph similar to Figure 4) but only the coefficients of the cut-off value of IMS ≥ 1.3 were statistically significant. This was expected as too few raters were classified as misfitting by IMS, compared to the OMS statistic (for that purpose, compare Figures 2 and 3). However, the pattern of the IMS results is clearly in the same direction as those of the OMS statistic.

Overall, our findings suggest that when the composition of the group of raters changes significantly, the group's equilibrium is destabilized and the probability of returning raters to be classified as misfitting is substantially higher. On the other hand, if the retention of raters is high, the probability of returning raters to be classified as misfitting diminishes even further, but the newcomers stick out with a high probability of being misfitting.

Interestingly, the total rating experience (either for the same or for other exams) is not significant, although the recent experience is important. Past experience (from more distant years) seems to have a rapidly diminishing effect. As mentioned before, with an average retention rate of 65%, the proportion of common raters between two non-directly adjacent years (e.g. year X and year X+2) is estimated to be around one third. Thus, a rater who rated in year X, skips X+1 and returns in year X+2, joins a rather unfamiliar community of raters.

In the context of the model of Table 3, the Intra-Class Correlation (ICC) coefficient is a measure of how well the repeated measures of the dependent variable within each rater resemble each other. In other words, we can practically interpret ICC as an indication of the overall differences between raters. The fact that we found intra-class correlations of 0.26 and 0.53 (Table 3; also see Appendices E and F) indicates a moderate degree of within-rater correlation of odds to be classified as misfitting. This suggests that being misfitting could be a personal characteristic to some degree, or could be related to some other background characteristics (e.g. personality). Unfortunately, we do not, currently, have the necessary information to investigate these speculations.

Is rating experience related to severity?

This section investigates the effect of rating experience on the rating characteristics across exams (RQ3). We investigated the effect of experience using all three different operationalizations of experience: (a) the cumulative experience with this particular exam, (b) the cumulative experience with similar exams, and (c) being a rater in the immediately preceding exam.

Table 4 presents a GLMM using rater Rasch severity estimates (in logits) as dependent variable. The variable "Experience" is a fixed effect (numeric covariate) and represents the number of years of experience with the specific exam. The model was fit with varying intercepts for raters in order to investigate the magnitude of the ICC and whether there was a significant between-rater variance of Rasch severity.

As shown in Table 4, the coefficient of Experience (experience with the same exam) is statistically significant and negative, albeit very small. This suggests that accumulating experience with this particular exam, may make the raters slightly less severe over time. The magnitude of the effect, however, is extremely small and practically negligible. When compared to the standard deviation of the examinee ability estimate, the effect of one additional year of experience on rater severity is around 1.6% of a standard deviation. In other words, six years of experience with this exam will only correspond to a reduction in severity equivalent to around one tenth of a standard deviation of the examinee ability distribution. There are very few raters with more than six years of experience, therefore the effect of this variable is practically negligible.

[Insert Table 4]

Being a rater in the immediately preceding exam did not yield statistically significant results. Experience with other exams was only marginally statistically significant but had a very small and practically negligible coefficient. Other variables, e.g., the proportion of returning raters per year, were also found not to have statistically significant coefficients and are not presented in the models.

Is it informative to use past data as a proxy of future severity and consistency??

This section investigates whether it is practically informative for researchers and practitioners to use past information as a proxy of the rating characteristics of the raters in subsequent exams (RQ4). Table 5 presents the Pearson correlations between the rating characteristics of the raters in consecutive exams. Column “r” represents the correlation between the Rasch severity estimates of the raters for two consecutive years; columns ‘Lower’ and ‘Upper’ represent the 95% lower and upper confidence intervals of the bootstrap resampling (1000 replications, percentile method). Statistically significant coefficients are indicated by a star (marginally non-significant coefficients are indicated by a cross). Column χ^2 represents the chi-square statistic of the crosstab between two dichotomous variables showing whether the same raters were classified as misfitting or not misfitting on two consecutive years, for an OMS ≥ 1.2 (other cut-off values gave similar results, see Appendix H). The last column represents the proportion of raters who rated in the current and the previous year.

[Insert Table 5]

For example, the third line of Table 5 demonstrates the correlations for the rating characteristics of the raters between the exams of years 2004 and 2005 (first column: “2004-2005”). The correlation between the Rasch measures of those 16 raters who rated both in 2004 and 2005 was 0.70 (the 95% bootstrap confidence interval suggests that the coefficient was statistically significant with a lower bound of 0.32 and an upper bound of 0.89). The chi-square test between their 2004 and 2005 classification as misfitting was 1.465 (p=0.226) when a cut-off OMS $\geq$ 1.2 is used.

To provide some extra information to the reader, Appendix H shows the same chi-square tests, calculated for all four cut-off values. The chi-square test between the 2004 and 2005 classification of raters as misfitting was 5.565 (p=0.018) for OMS $\geq$ 1.15. For the 1.2 and 1.25 cut-off values the results were not statistically significant.

Overall, Table 5 and Appendix H can be very useful because they illustrate the degree to which one might aspire to use past data as a proxy of the future rating behaviour of the raters. It is interesting to see a high concordance from year to year on several occasions, especially when the number of returning raters is high. However, we would need significantly more research until we reach the point of being able to make practically useful predictions.

Discussion and Recommendations

Drawing on Knoch & Chapelle's (2018) conceptualizations of the rating process we conclude that the evaluation inferences of the validity argument of the EFL exam under study cannot be fully supported. For example, the rating-relevant warrant underlining the evaluation inference of this study, which refers to whether raters are able to rate test tasks reliably, was not adequately supported by our findings. The results of the Many-facet Rasch showed that raters could not apply the scale consistently to each test task. Moreover, we found that measurable rater characteristics (such as participating in the previous examination as a rater) was a significant determinant of their probability to be inconsistent. This finding is very important as it weakens the validity argument of the examination and could undermine the trust of stakeholders. However, the substantial instability in the rating characteristic of individual raters across exams is not a unique characteristic of this particular examination (see also Author, 2006, 2018; Congdon & McQueen, 2000; Huhta et al., 2014; Leckie & Baird, 2011; Wolfe, 2009).

Other than the above, the study breaks new ground by building on the existing literature operationalizing “experience” in multiple ways. First of all, we used three different measures of experience: (a) cumulative experience in the same exam, (b) cumulative experience in different exam, and (c) recent experience. We have found that first two had practically insignificant effects on rating characteristics. Accumulating experience in other (albeit similar) language exams did not have any transferable effect on rating in this particular exam. This, in effect, supports the findings of Huhta et al. (2014) who claim that raters often have difficulty in working with the rating scales: each rating scale is different and experience with different types of rating scales is not necessarily transferable.

However, recent experience is very important. It was found that, being a member of the CoP in the immediately preceding exam is a significant determinant of rating characteristics (also supported by Myford and Wolfe, 2009). For example, the probability of returning raters to be inconsistent is almost halved (from around 27% to than 17%), as the proportion of returning raters doubles (from 0.4 to 0.8). Also, the probability of non-returning raters to be classified as inconsistent increases by a factor of more than six (from around 8% to almost 50%) as the proportion of returning raters doubles (from 0.4 to 0.8). Our findings are reasonable: tests, regulations, student populations and other important factors may change slightly across years. If experience is meant to have some effect, it will be the most recent experience, not the distant one, whose effects seem to diminish quite rapidly. This conclusion, provides insightful information since longitudinal studies in rater development usually involve only a temporal dimension (Lim, 2011).

It was also encouraging that we found statistically significant and sizeable correlations between the rating characteristics of raters from year to year on some occasions. Although we failed to find statistically significant correlations across all years, this might be because of several factors such as the number of raters and small curriculum changes across time, which may render experience obsolete (see the Methods section). The magnitude of the correlations was large in almost all of the cases, even if statistical significance was not typically reached. This is an indication that, with more data, and probably with the help of further research, exam boards might be able to use past data as proxies of future rating characteristics.

Finally, because of the design of our study, we were – for the first time – able to investigate quantitatively the concept of the "Community of Practice". We have shown that the CoP sets the standards through consensus and those who are more likely to be classified as misfitting are the newcomers, especially if the community has retained a large proportion of its past membership. On the other hand, even the most experienced raters are likely to be classified as misfitting, if the membership of the community changes dramatically and the equilibrium of the community gets destabilized.

Due to our experience with the current study, we are in a position to propose specific methodological recommendations for future research. Firstly, we would like to engage critically with existing research and suggest that, to monitor stability across time effectively, it is useful to invest in longitudinal designs which span across substantial periods of time (e.g. many months or years). Shorter designs (e.g. less than a month) can be very useful for specific research questions (e.g. when investigating rating fatigue), but can also be influenced by the spontaneity of the moment and other random variations (e.g. temporary mood). In the field of language assessment, short term designs are more likely to have "thin" data points (fewer observations per time unit) as it is difficult for raters to rate many scripts within a short period of time. When planning for short term designs, researchers must make sure that they will have enough datapoints per time unit (e.g. per day).

In relation to the design of the existing studies, although the contribution of experiments in the literature is important, studies using operational data may enjoy a high degree of external validity. For example, we were impressed by the low retention of raters in our study but also by their puzzling patterns of participation. Due to our close collaboration with the exam body that provided the data, it was interesting to observe how raters often disappeared for a year or two and then reappeared as "experienced" raters. Building on the pioneer work of Lim (2011), our study encourages the research community to invest on multi-year longitudinal designs based on operational tests.

A word of caution relating to the operational definition of some key variables is probably long over-due. Our close engagement with the logistics of the operational rating and our observation that raters disappear and re-appear almost randomly, motivated multiple operational definitions of the key variable of "experience". Thus, we used (a) a variable for the experience in this exam accumulated across time, (b) a variable of recent experience (participation in the immediately previous exam) and (c) a variable for the accumulated experience in other, similar, language exams. We have shown that these are not only theoretically, but also empirically, distinct variables and should be treated as such in future studies because they affect rating characteristics in different ways.

Finally, we would also like to make some recommendations for the practitioners. Our research has shown that experience with the same or other, similar, exams, does not necessarily guarantee desirable rating characteristics. What is of great importance is how well a rater has integrated within the CoP. Participation in the immediately preceding exam could be used as a crude proxy of this integration and these raters should have a hiring priority.

Also, it is important to offer incentives to raters to participate in subsequent rating cycles without skipping any of them. This is especially important in contexts where there is low retention rate. It is also important to develop efficient techniques to help newcomers increase their sense of identity and community belonging, so as to increase their likelihood for retention.

Finally, we feel that a qualitative study could have contributed significantly to the understanding of the micro-mechanisms that govern the stability of the community of practice. A qualitative study could have helped us to investigate how group dynamics govern the negotiation of standards and how these are re-negotiated in times of radical change. This knowledge could be used to develop tools for rater training and to provide constructive feedback to raters.

References

Author1 & Colleague (2004)

Author1 (2006)

Author1 (2018)

Author1 (2020).

Colleagues and Author1 (2018)

Al-Maamari, M. (2016). Community of assessment practice or interests: The case of EAP writing assessment. Indonesian Journal of Applied Linguistics, 5(2), 272-281. https://doi.org/ 10.17509/ijal.v5i2.1351

Barkaoui, K. (2010a). Explaining ESL essay holistic scores: A multilevel modelling approach. *Language Testing*, *27*(4), 515-535. https://doi.org/10.1177/0265532210368717

Barkaoui, K. (2010b). Variability in ESL Essay Rating Processes: The Role of the Rating Scale and Rater Experience. *Language Assessment Quarterly*, 7, 54–74. https://doi.org/10.1080/15434300903464418

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49-58. Available at http://iej.cjb.net HSC marking. (2019, 14 February). Available at www.boardofstudies.nsw.edu.au/hsc_exams/marking.html

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

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

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

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, *32*(4), 521–541. https://doi.org/10.1177/0265532215575626

Douglas, S. R. (2015). The Relationship Between Lexical Frequency Profiling Measures and Rater Judgements of Spoken and Written General English Language Proficiency on the CELPIP-General Test. *TESL Canada Journal*, *32*(9), 43–64. https://doi.org/10.18806/tesl.v32i0.1217

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. Language Testing, *24*(1), 37–64. https://doi.org/10.1177/0265532207071511

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, *11*(2), 195-208.https://doi.org/10.1207/s15324818ame1102_5

Gianinazzi, M. E., Rueegg, C. S., Zimmerman, K., Kuehni, C. E., Michel, G., & the Swiss Paediatric Oncology Group (SPOG). (2015). Intra-Rater and Inter-Rater Reliability of a Medical Record Abstraction Study on Transition of Care after Childhood Cancer. *PLOS ONE*, *10*(5), 124-290. https://doi.org/10.1371/journal.pone.0124290

Haj-Ali, R., &Feil, P. (2006). Rater reliability: short- and long-term effects of calibration training. *Journal of Dentistry Education*, *70*(4), 428-433. Available at https://www.ncbi.nlm.nih.gov/pubmed/16595535

Han, C. (2015). Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach. *Interpreting*, *17*(2), 255–283. https://doi.org/10.1075/intp.17.2.05han

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R.J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43-58. https://doi.org/10.1111/j.1745-3984.2009.01068.x

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.

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

Isbell, D.R. (2017). Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. Assessing Writing, 34, 37-49. https://doi.org/10.1016/j.asw.2017.08.004

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

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

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31. https://doi.org/10.1191/0265532202lt218oa

Kuiken, F., & Vedder, I. (2014). Raters' decisions, rating procedures and rating scales. *Language Testing*, *31*(3), 279–284. https://doi.org/10.1177/0265532214526179

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477–499. https://doi.org/10.1177/0265532217710049

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

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* (Doctor of Philosophy). University of Michigan, USA.

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*. Chicago: MESA Press.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2005). A user's guide to FACETS: Rasch-model computer programs [Software manual]. Chicago, IL: Winsteps.com.

MATSEC Support Unit. (2018). *Paper Setting: Procedures and Good Practices*. Malta: Università ta' Malta.

Myford, C., & 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. (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

NSW Government. (2019). Home | NSW Education Standards. Retrieved 10 January 2019, from https://www.educationstandards.nsw.edu.au/wps/portal/nesa/home

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, *26*(2), 161-186. https://doi.org/10.1177/0265532208101005

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. Assessing Writing, 31, 98–112. https://doi.org/10.1016/j.asw.2016.08.005

Shay, S. (2005). The assessment of complex tasks: a double reading. Studies in Higher Education, 30(6), 663–679. https://doi.org/10.1080/03075070500339988

Slomp, D. & East, M. (Eds.). (2019). Framing the Future of Writing Assessment [Special issue]. Assessing Writing, 42.

Wang, J., & Engelhard, G. (2019). Exploring the Impersonal Judgments and Personal Preferences of Raters in Rater-Mediated Assessments With Unfolding Models. *Educational and Psychological Measurement*, *79*(4), 773-795. https://doi.org/10.1177/0013164419827345

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

Wenger, E., McDermott, R., & Snyder, W. M. (2002). Cultivating communities of practice. Boston, MASS, Harvard Business School Press.

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

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 38–54. https://doi.org/10.1016/j.asw.2015.05.002

Wolfe, E. W. (2009). Item and rater analysis of constructed response items via the multi-faceted Rasch model. *Journal of Applied Measurement*, *10*(3), 335-347.

Zhang, B., Chen, Z., &Albert, P. S. (2012). Estimating diagnostic accuracy of raters without an old standard by exploiting a group of experts. *Biometrics, 68*(4), 1294–1302. https://dx.doi.org/10.1111%2Fj.1541-0420.2012.01789.x

Table 1

Frequency of Rater Participation in the Thirteen Examinations

Frequency of participation as a rater	1	2	3	4	5	6	7	8	9	10	11	12
Number of raters	39	19	8	11	7	4	3	1	3	1	2	1

Table 2

Number of raters per year, mean of experience, and proportion of common raters for consecutive years

Exam Years	Current year (N of raters)	Mean Experience	Previous year (N of raters)	Proportion who rated in previous year
2002-2003	24	1.46	11	0.46
2003-2004	25	2.04	17	0.68
2004-2005	19	2.89	16	0.84
2005-2006	28	2.18	11	0.39
2006-2007	20	3.25	15	0.75
2007-2008	24	3.38	15	0.63
2008-2009	21	3.81	16	0.76
2009-2010	21	3.43	14	0.67
2010-2011	27	4.22	18	0.67
2011-2012	26	4.73	20	0.77
2012-2013	23	4.57	14	0.61
2013-2014	25	4.00	13	0.52
Mean	23.58	3.33	15.00	0.65
SD	2.84	1.03	2.66	0.13

Table 3

Determinants of inconsistency (Outfit Mean Square cut-off value ≥1.2)

Dependent Variable: Rater not Misfitting = 0 / Rater misfitting = 1

	Outfit MS ≥ 1.2			**Infit MS ≥ 1.2**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.01	0.00 – 0.16	**0.001**	0.02	0.00 – 0.53	**0.021**
Proportion returning	263.20	3.89 – 17829.31	**0.010**	22.97	0.13 – 3929.47	0.232
Individual rater returning	56.71	1.70 – 1893.95	**0.024**	16.76	0.15 – 1820.15	0.239
Proportion returning X Individual returning	≈ 0.01	0.00 – 0.22	**0.012**	≈ 0.01	0.00 – 5.15	0.128
Random Effects						
σ^2	3.29			3.29		
τ_{00}	1.16 $_{\text{Rater}}$			3.65 $_{\text{Rater}}$		
ICC	0.26			0.53		
N	89 $_{\text{Rater}}$			89 $_{\text{Rater}}$		
Observations	283			283		
Marginal R^2 / Conditional R^2	0.047 / 0.297			0.033 / 0.542		

Table 4

Determinants of Rasch severity (Intercepts as Random Effects and Experience as a Fixed Effect)

	Rasch severity (in logits)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-0.02	-0.14 – 0.10	0.745
Experience	-0.05	-0.09 – -0.01	**0.021**
Random Effects			
σ^2	0.44		
$\tau_{00\ \text{marker}}$	0.13		
ICC	0.22		
N $_{\text{marker}}$	89		
Observations	283		
Marginal R^2 / Conditional R^2	0.024 / 0.242		

Table 5

Year to year correlations of rater characteristics

	Rasch Severity			**Misfit** **(Outfit ≥1.2)**	**Previous year** **(N of raters)**	**Proportion who rated** **in previous year**
Exam Years	r	Lower	χ^2	χ^2		
2002-2003	0.55	-0.07	0.87	NA	11	0.46
2003-2004	**0.62***	0.19	0.85	χ^2(1)=0.142, p=0.706	17	0.68
2004-2005	**0.70***	0.32	0.89	χ^2(1)=1.465, p=0.226	16	0.84
2005-2006	0.55	-0.07	0.86	χ^2(1)=1.925, p=0.165	11	0.39
2006-2007	0.27	-0.28	0.69	χ^2(1)=2.50, p=0.114	15	0.75
2007-2008	0.19	-0.36	0.64	χ^2(1)=0.001, p=0.999	15	0.63
2008-2009	0.45	-0.05	0.78	χ^2(1)=0.872, p=0.350	16	0.76
2009-2010	0.16	-0.41	0.63	χ^2(1)=2.363, p=0.124	14	0.67
2010-2011	0.19	-0.33	0.58	χ^2(1)=3.583, p=0.058$^+$	18	0.67
2011-2012	-0.29	-0.68	0.10	χ^2(1)=8.235, **p=0.004***	20	0.77
2012-2013	0.51	-0.01	0.82	χ^2(1)=0.636, p=0.425	14	0.61
2013-2014	0.38	-0.22	0.78	χ^2(1)=5.318, **p=0.021***	13	0.52

Note:

+ indicates a marginally non-significant test at the 0.05 level

* indicates a significant test at the 0.05 level

NA indicates that a chi-square test could not be calculated (e.g. for year 2002, there were no raters with fit statistics larger than 1.2, so the variable consisted only of 0's and had no 1's)

Appendix A

An example of the structure of the EFL examination of the University Entrance Exam (for the years 2002-2011)

Section	Duration	Description
Writing	**2 hours 30 min**	***Task 1:*** 250-300 words of continuous prose or description or argument in response to a short stimulus **(25 points)**
Reading		***1 reading text*** ***Task 1*:** Multiple-choice questions testing skim-/gist-reading skills (10 points) ***Task 2*:** Three open-ended questions testing more detailed comprehension (9 points). ***Task 3*:** Extended writing 80–100 words in response to two short stimuli questions based on the text and students' personal opinion (15 points) **Task 4:** Multiple-matching questions testing understanding of unknown vocabulary in the text (6 points) **(40 points)**
Language Usage		*(testing grammar, syntax and vocabulary)* ***Task 1*:** Sentence transformations (5 points) ***Task 2*:** Cloze passage (5 points). ***Task 3*:** Modified Cloze passage (5 points). **(15 points)**
Listening	**Approx. 15 min**	A monologue (one person speaking), or a recording with two or more speakers lasting approximately 3-4 minutes. ***Task 1*:** Multiple-choice questions (5 points) ***Task 2*:** True/False questions (5 points). ***Task 3*:** Modified Cloze summarising the oral input (10 points) **(20 points)**

Note: The structure of the examination changed several times.

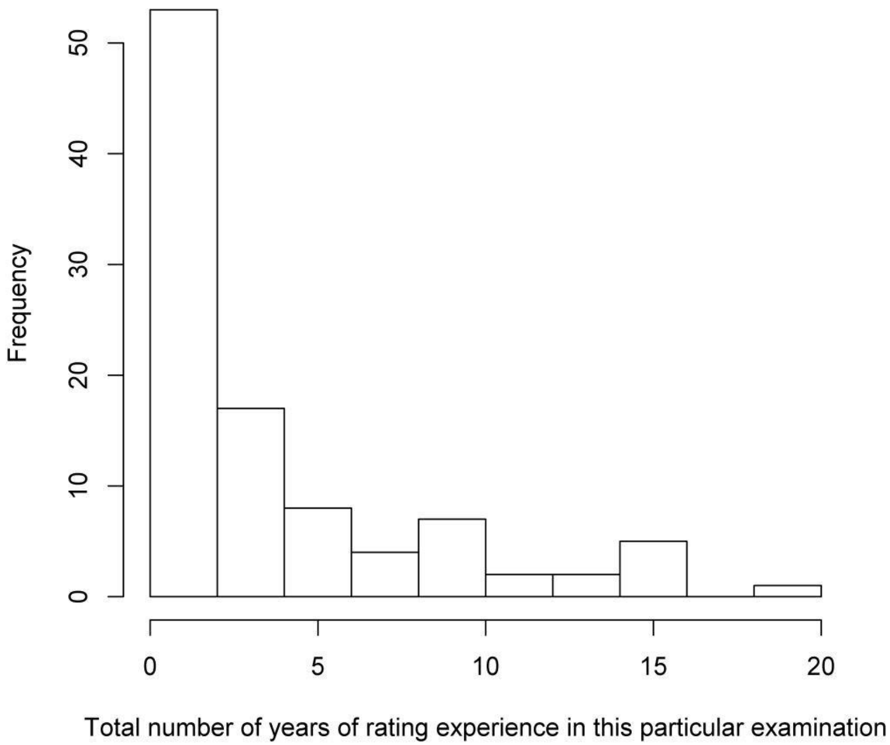
Appendix B

A sample of the rating scale for the Writing component of the EFL examination

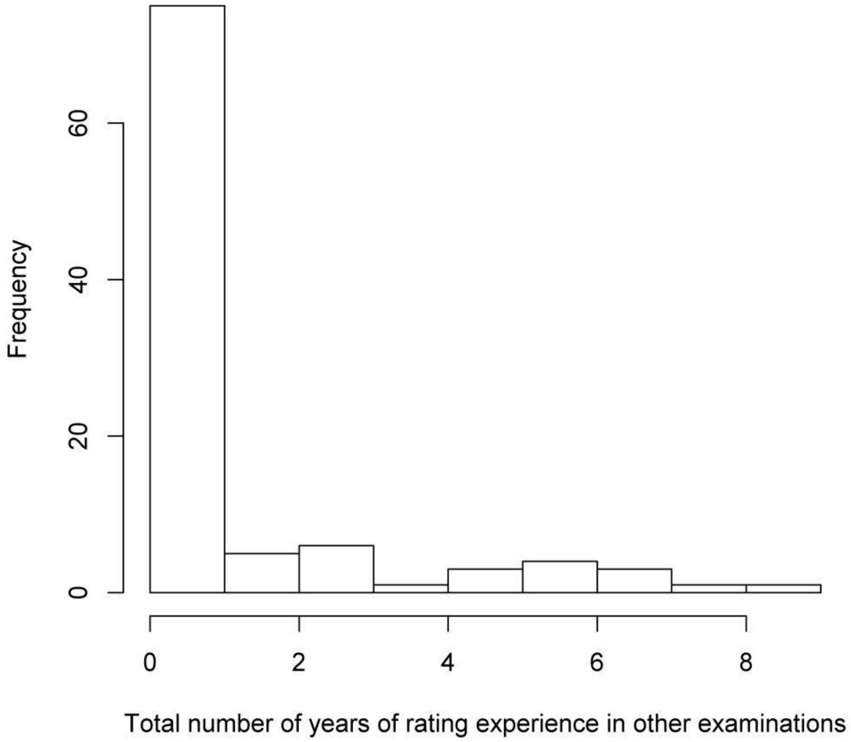
Criterion	Descriptors
A. Content **(8-6-3-1)**	**8 -** Relevant to assigned topic. Thorough development of topic. **6 -** Mostly relevant to topic. Limited development. **3 -** Slightly relevant to topic. Inadequate development of topic. **1** -Almost no relevance to topic.
B. Organisation **(8-6-3-1)**	**8** – Ideas clearly stated and supported. Logical sequencing and cohesion. **6-** Main ideas stand out but not fully supported. Logical but incomplete sequencing. **3** – Ideas confused and/or disconnected. Lacks logical sequencing and development. Some paragraphs well-constructed. **1** – No organization. Paragraphs almost non-existent.
C. Vocabulary **(6-4-2-1)**	**6 –** Correct word/idiom choice and usage. Extensive range of vocabulary. **4-** Occasional errors of word/idiom choice and usage but meaning not obscured. **2-** Frequent errors of word/idiom choice and usage. Meaning confused or obscured. **1-** Very little use of vocabulary, idioms, word form.
D. Grammar **(5-3-1)**	**5-** Few errors of agreement, tense, word order, article, pronouns, prepositions. **3-** Frequent errors of agreement, tenses, word order, articles, pronouns, prepositions. **1** – Dominated by errors.
E. Mechanics **(3-1)**	**3** – Few errors of spelling and punctuation. **1** – Dominated by errors of spelling and punctuation.
When ideas and information bear no resemblance to the topic, the composition receives no marks	

Note: The structure of the scale changed several times (the sample above corresponds to the years 2012-2014). The rating scale consisted of a number of criteria (e.g. for the period 2012-2014, the scale had five criteria). Each of the criteria was treated as a different polychotomous "item" in the Rasch analysis. In the case where a criterion was slightly modified (e.g. having five instead of four levels of performance), this was treated as a new, different, "item" and the "old item" was treated as structurally missing data in future examination years.

Appendix C



Mean: 4.2, Median: 2.0, Standard Deviation: 4.3.



Mean: 1.1, Median: 0, Standard Deviation: 2.2.

Appendix D

Model-data Fit for Items, Examinees and Raters of the Thirteen Datasets

Month	Exam	Items		Examinees		Raters		Separation Index (reliability) sample size (a.s.e.)	
		Infit	Outfit	Infit	Outfit	Infit	Outfit	Candidates	Raters
Baseline 0	June 2002	0.97* (0.42)	1.00 (0.46)	0.90 (0.78)	1.00 (0.91)	0.90 (0.12)	1.00 (0.13)	3.73 (0.93) N=3237	11.40 (0.99), N=21, (a.s.e=0.03)
12	June 2003	1.01 (0.41)	1.02 (0.44)	0.94 (0.83)	1.02 (0.94)	0.94 (0.12)	1.02 (0.12)	4.07 (0.94) N=2739	10.14 (0.99), N=24, (a.s.e=0.04)
24	June 2004	1.00 (0.47)	1.01 (0.49)	0.89 (0.79)	1.01 (0.96)	0.89 (0.18)	1.02 (0.17)	3.69 (0.93) N=2438	9.01 (0.99), N=25, (a.s.e=0.04)
36	June 2005	0.99 (0.42)	0.98 (0.40)	0.85 (0.74)	0.96 (0.84)	1.21 (0.12)	1.03 (0.10)	3.32 (0.92) N=1700	10.09 (0.99), N=19, (a.s.e=0.04)
48	June 2006	1.05 (0.27)	2.32 (2.62)	1.00 (0.64)	1.23 (1.56)	1.01 (0.21)	1.64 (1.61)	3.72 (0.93) N=1975	10.69 (0.99), N=28, (a.s.e=0.05)
60	June 2007	1.00 (0.16)	1.15 (0.40)	0.96 (0.74)	1.06 (1.09)	1.02 (0.24)	1.18 (0.33)	5.51 (0.97) N=1640	16.56 (0.99), N=20, (a.s.e=0.06)
72	June 2008	1.01 (0.15)	1.07 (0.26)	0.96 (0.73)	1.02 (0.93)	1.01 (0.22)	1.07 (0.35)	4.26 (0.95) N=1672	9.92 (0.99), N=24, (a.s.e=0.06)
84	June 2009	1.01 (0.21)	1.24 (0.62)	0.94 (0.66)	1.10 (1.27)	0.97 (0.19)	1.23 (0.69)	5.19 (0.96) N=1235	11.84 (0.99), N=21, (a.s.e=0.06)
96	June 2010	1.00 (0.15)	1.22 (0.55)	0.95 (0.72)	1.06 (1.19)	0.98 (0.25)	1.23 (0.58)	5.03 (0.96) N=1608	10.81 (0.99), N=21, (a.s.e=0.06)
108	June 2011	1.01 (0.14)	1.26 (0.60)	0.99 (0.69)	1.03 (1.04)	1.00 (0.26)	1.30 (1.15)	5.24 (0.96) M=1789	8.85 (0.99), N=27, (a.s.e=0.07)
120	June 2012	1.00 (0.14)	1.04 (0.19)	0.98 (0.33)	1.05 (0.34)	0.95 (0.65)	1.03 (0.77)	5.27 (0.97) N=1779	12.36 (0.99), N=26, (a.s.e=0.06)
132	June 2013	1.00 (0.12)	1.02 (0.16)	0.95 (0.80)	0.99 (0.83)	1.00 (0.26)	1.02 (0.26)	4.09 (0.94) N=1773	13.53 (0.99), N=23, (a.s.e=0.05)
144	June 2014	1.00 (0.13)	1.05 (0.19)	0.98 (0.65)	1.04 (0.81)	0.99 (0.22)	1.06 (0.30)	5.70 (0.97) N=1709	12.94 (0.99), N=25, (a.s.e=0.06)

Note. Average values and standard deviations (in parentheses). N indicates the sample size. Finally, ‘a.s.e’. stands for “average standard error” and is simply the average of the standard error of the rater Rasch measures.

Appendix E

Determinants of misfit (OMS)

Dependent Variable: Rater Not Misfitting = 0 / Rater Misfitting = 1, using four OMS cut-off values

	Outfit MS ≥ 1.15			**Outfit MS ≥ 1.2**			**Outfit MS ≥ 1.25**			**Outfit MS ≥ 1.3**		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.01	0.00 – 0.16	**0.001**	0.01	0.00 – 0.16	**0.001**	0.02	0.00 – 0.23	**0.003**	0.00	0.00 – 0.05	**<0.001**
Proportion returning	421.60	7.33 – 24249.61	**0.003**	263.20	3.89 – 17829.31	**0.010**	93.43	1.33 – 6555.24	**0.036**	2051.14	9.40 – 447623.41	**0.006**
Individual rater returning	187.82	6.58 – 5361.51	**0.002**	56.71	1.70 – 1893.95	**0.024**	27.50	0.77 – 981.71	0.069	119.58	1.39 – 10270.73	**0.035**
Proportion returning X Individual returning	≈ 0.01	0.00 – 0.02	**0.001**	≈ 0.01	0.00 – 0.22	**0.012**	≈ 0.01	0.00 – 0.84	**0.043**	≈ 0.01	0.00 – 0.32	**0.022**
Random Effects												
σ^2	3.29			3.29			3.29			3.29		
τ_{00}	1.02 $_{\text{Rater}}$			1.16 $_{\text{Rater}}$			1.09 $_{\text{Rater}}$			1.52 $_{\text{Rater}}$		
ICC	0.24			0.26			0.25			0.32		
N	89 $_{\text{Rater}}$			89 $_{\text{Rater}}$			89 $_{\text{Rater}}$			89 $_{\text{Rater}}$		
Observations	283			283			283			283		
Marginal R^2 / Conditional R^2	0.070 / 0.291			0.047 / 0.297			0.033 / 0.273			0.072 / 0.365		

Appendix F

Determinants of misfit (IMS)

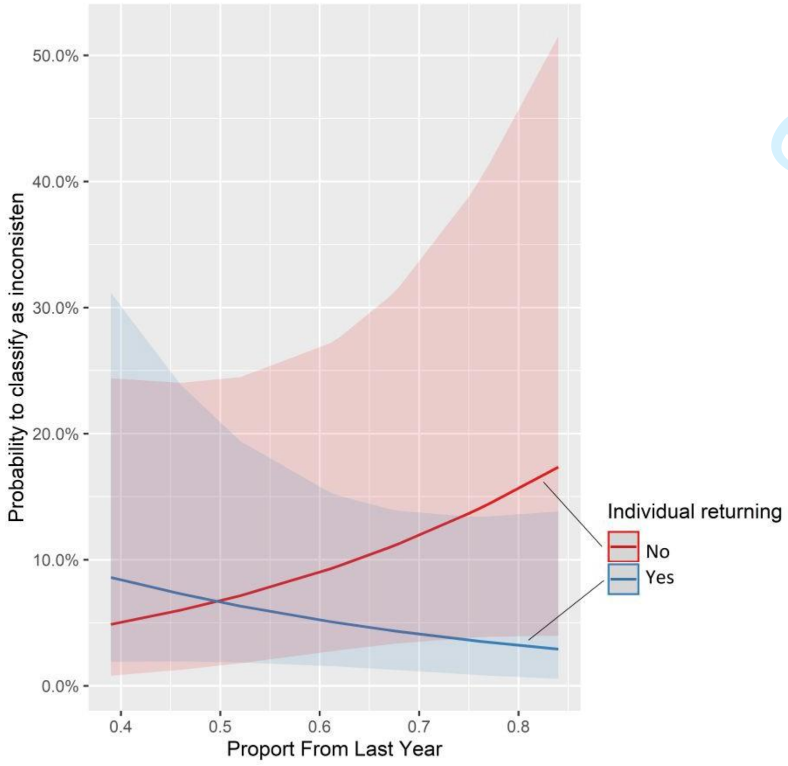
Dependent Variable: Rater Not Misfitting = 0 / Rater Misfitting = 1, using four IMS cut-off values

	Infit MS ≥ 1.15			Infit MS ≥ 1.2			Infit MS ≥ 1.25			Infit MS ≥ 1.3		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	0.09	0.01 – 1.27	0.075	0.02	0.00 – 0.53	**0.021**	0.01	0.00 – 0.44	**0.018**	≈ 0.01	0.00 – 0.01	**0.001**
Proportion returning	3.77	0.06 – 256.87	0.538	22.97	0.13 – 3929.47	0.232	43.21	0.20 – 9351.05	0.170	1506.09	0.24 – 928863.67	0.100
Individual rater returning	3.36	0.08 – 149.19	0.532	16.76	0.15 – 1820.15	0.239	15.84	0.10 – 2511.55	0.285	976.94	0.50 – 191295.55	0.075
Proportion returning X Individual returning	0.05	0.00 – 18.46	0.319	≈ 0.01	0.00 – 5.15	0.128	≈ 0.01	0.00 – 6.57	0.135	≈ 0.01	0.00 – 0.69	**0.043**
Random Effects												
σ^2	3.29			3.29			3.29			3.29		
τ_{00}	1.73 $_{\text{Rater}}$			3.65 $_{\text{Rater}}$			3.91 $_{\text{Rater}}$			64.15 $_{\text{Rater}}$		
ICC	0.34			0.53			0.54			0.95		
N	89 $_{\text{Rater}}$			89 $_{\text{Rater}}$			89 $_{\text{Rater}}$			89 $_{\text{Rater}}$		
Observations	283			283			283			283		
Marginal R^2 / Conditional R^2	0.029 / 0.364			0.033 / 0.542			0.047 / 0.565			0.010 / 0.952		

Appendix G

Predicted probabilities for a rater to be classified as misfitting (IMS ≥ 1.2), for different values of the proportion of returning raters.

The figure corresponds to the second model of Appendix F (i.e., the "Infit MS ≥ 1.2" model of Table 3). We observe the same patterns as in Figure 3 (i.e., the "Outfit MS ≥ 1.2" model of Table 3) but the coefficients are not statistically significant.



Appendix H

Year to year correlations of rater characteristics

	Rasch Severity			Misfit (Outfit ≥1.15)	Misfit (Outfit ≥1.2)	Misfit (Outfit ≥1.25)	Misfit (Outfit ≥1.3)	Previous year (N of raters)	Proportion who rated in previous year
Exam Years	r	Lower	χ^2	χ^2	χ^2	χ^2	χ^2		
2002-2003	0.55	-0.07	0.87	$\chi^2(1)=2.93$, p=0.087	NA	NA	NA	11	0.46
2003-2004	0.62*	0.19	0.85	$\chi^2(1)=0.781$, p=0.377	$\chi^2(1)=0.142$, p=0.706	NA	NA	17	0.68
2004-2005	0.70*	0.32	0.89	$\chi^2(1)=5.565$, p=0.018*	$\chi^2(1)=1.465$, p=0.226	$\chi^2(1)=0.152$, p=0.697	$\chi^2(1)=0.001$, p=0.999	16	0.84
2005-2006	0.55	-0.07	0.86	$\chi^2(1)=0.917$, p=0. 338	$\chi^2(1)=1.925$, p=0.165	NA	NA	11	0.39
2006-2007	0.27	-0.28	0.69	$\chi^2(1)=5.00$, p=0.025*	$\chi^2(1)=2.50$, p=0.114	$\chi^2(1)=2.50$, p=0.114	$\chi^2(1)=1.153$, p=0.282	15	0.75
2007-2008	0.19	-0.36	0.64	$\chi^2(1)=1.111$, p=0.298	$\chi^2(1)=0.001$, p=0.999	$\chi^2(1)=0.085$, p=0.770	$\chi^2(1)=0.416$, p=0.519	15	0.63
2008-2009	0.45	-0.05	0.78	$\chi^2(1)=2.798$, p=0.094	$\chi^2(1)=0.872$, p=0.350	$\chi^2(1)=3.419$, p=0.065$^+$	$\chi^2(1)=3.419$, p=0.064$^+$	16	0.76
2009-2010	0.16	-0.41	0.63	$\chi2(1)=1.167$, p=0.280	$\chi^2(1)=2.363$, p=0.124	$\chi^2(1)=3.764$, p=0.052$^+$	$\chi^2(1)=5.915$, p=0.015*	14	0.67
2010-2011	0.19	-0.33	0.58	$\chi^2(1)=4.923$, p=0.026*	$\chi^2(1)=3.583$, p=0.058$^+$	$\chi^2(1)=5.716$, p=0.017*	$\chi^2(1)=5.716$, p=0.017*	18	0.67
2011-2012	-0.29	-0.68	0.10	$\chi^2(1)=8.235$, p=0.004*	$\chi^2(1)=8.235$, p=0.004*	$\chi^2(1)=8.235$, p=0.004*	$\chi^2(1)=8.235$, p=0.004*	20	0.77

2012-2013	0.51	-0.01	0.82	$\chi^2(1)=0.321$, p=0.571	$\chi^2(1)=0.636$, p=0.425	$\chi^2(1)=0.636$, p=0.425	$\chi^2(1)=0.636$, p=0.425	14	0.61
2013-2014	0.38	-0.22	0.78	$\chi^2(1)=1.00$, p=0.317	$\chi^2(1)=5.318$, p=0.021*	$\chi^2(1)=0.965$, p=0.326	$\chi^2(1)=0.430$, p=0.512	13	0.52

Note:

+ indicates a marginally non-significant test at the 0.05 level

* indicates a significant test at the 0.05 or lower level

NA indicates that a chi-square test could not be calculated (e.g. for year 2002, there were no raters with fit statistics larger than 1.2, so the variable consisted only of 0's and had no 1's)

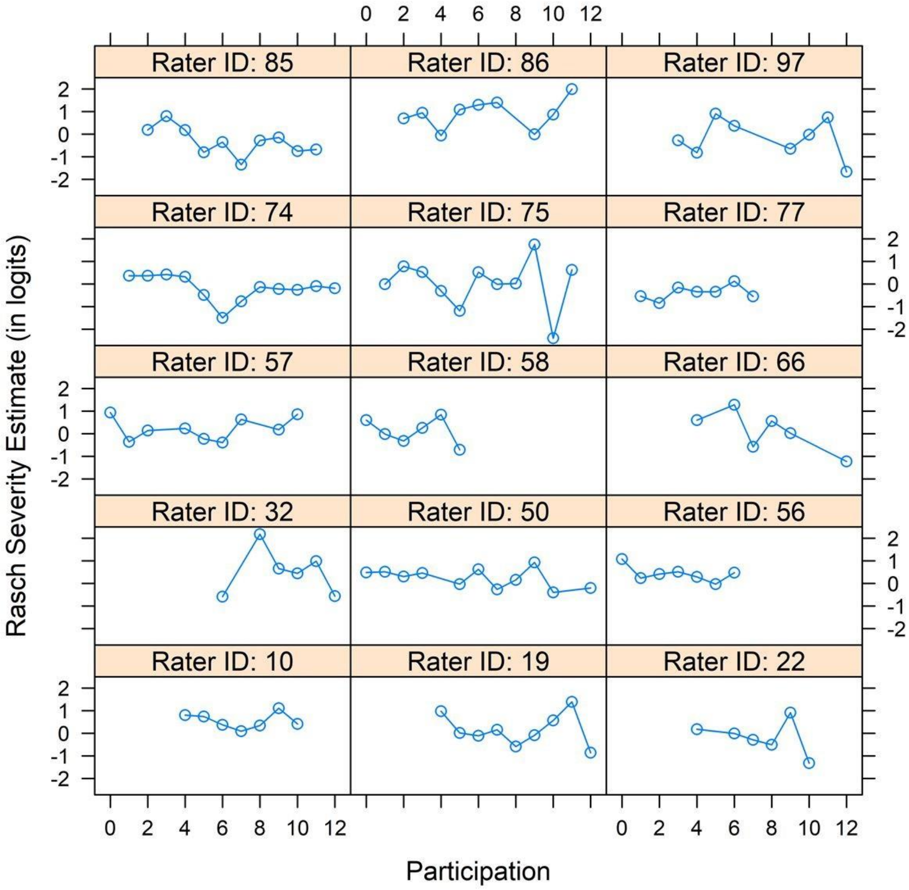


Figure 1. Rasch severity estimates, for each rater, for each examination (only for raters who participated in six or more examinations). The x-axis ranges from 0 (the baseline year = Year 2002) to 12 (the last data set = Year 2014).

150x142mm (216 x 216 DPI)

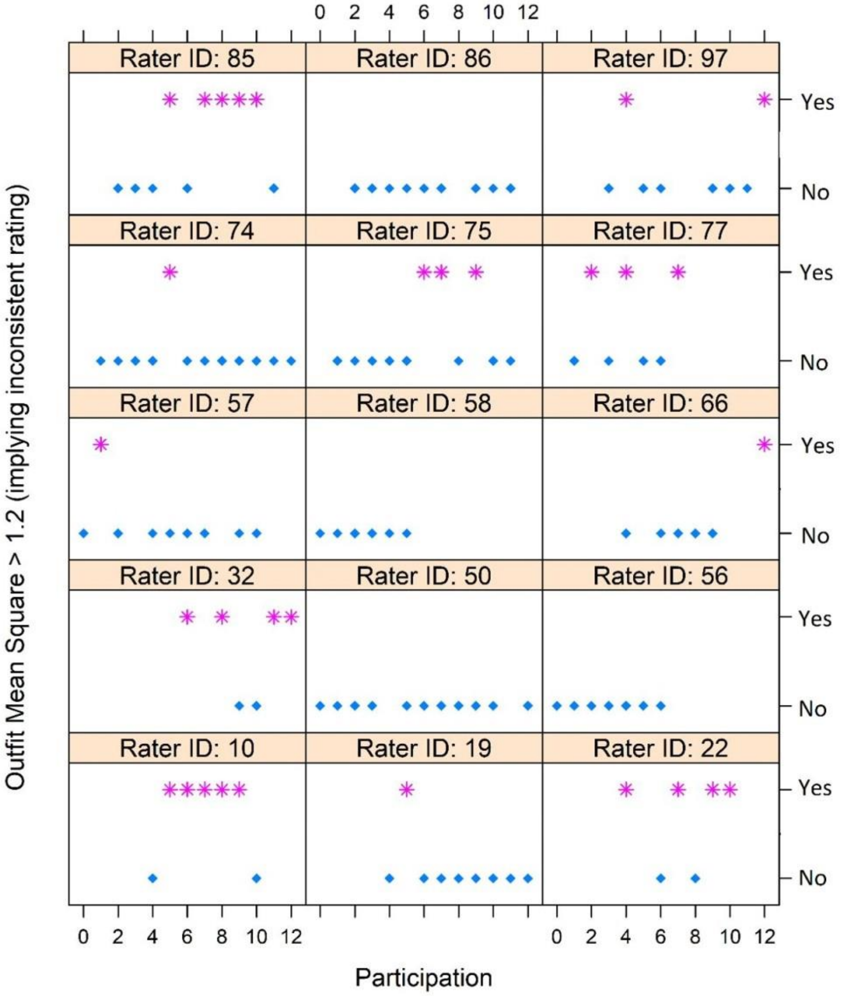


Figure 2. Outfit Mean Square statistics, for each rater, for each examination. To improve readability, the figure only shows raters who participated in six or more examinations. An asterisk shows raters with fit statistics larger than the rule of thumb (Outfit Mean Square ≥ 1.2 for a particular year). The x-axis ranges from 0 (the baseline year = Year 2002) to 12 (the last data set = Year 2014).

168x169mm (216 x 216 DPI)

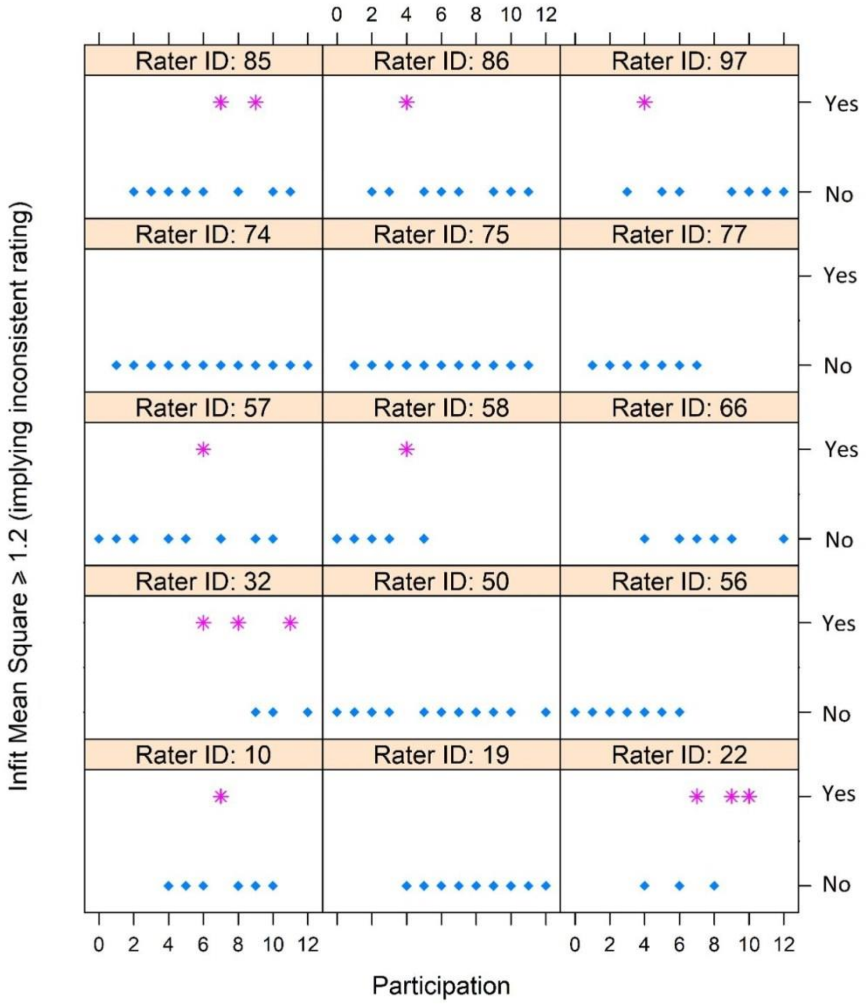


Figure 3. Infit Mean Square statistics, for each rater, for each examination. To improve readability, the figure only shows raters who participated in six or more examinations. An asterisk shows raters with fit statistics larger than the rule of thumb (Infit Mean Square ≥ 1.2 for a particular year). The x-axis ranges from 0 (the baseline year = Year 2002) to 12 (the last data set = Year 2014).

167x172mm (216 x 216 DPI)

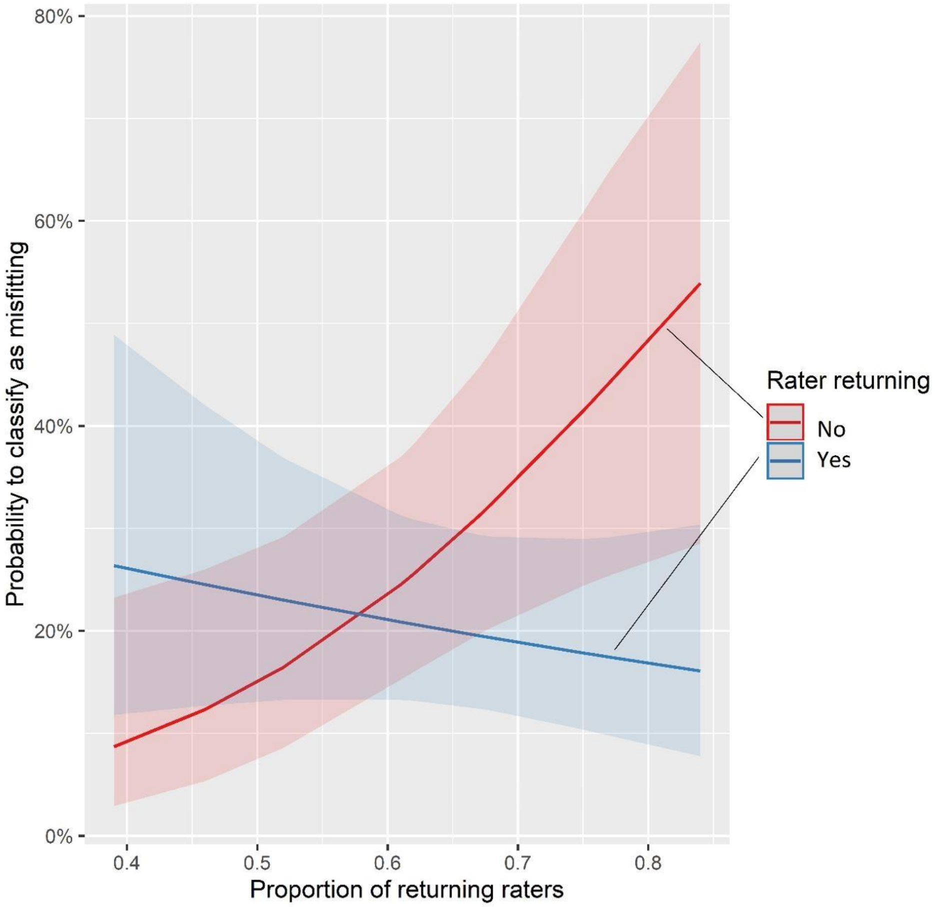


Figure 4. Predicted probabilities for a rater to be classified as misfitting (OMS ≥ 1.2), for different values of the proportion of returning raters.

165x160mm (216 x 216 DPI)