Psychological Test and Assessment Modeling, Volume 63, 2021 (1), 65–94

Measuring Rater Centrality Effects in Writing Assessment: A Bayesian Facets Modeling Approach

Thomas Eckes[1] *& Kuan-Yu Jin*[2]

Abstract

Rater effects such as severity/leniency and centrality/extremity have long been a concern for researchers and practitioners involving human raters in performance assessments. In the present research, a facets modeling approach advanced by Jin and Wang (2018) was adopted to account for both rater severity and centrality effects in a writing assessment context. In two separate studies, raters scored examinees' writing performances on a set of criteria using a four-category rating scale. Rater severity and centrality parameters were estimated building on Bayesian Markov chain Monte Carlo methods implemented in the freeware JAGS run from within the R environment. The findings revealed that (a) raters differed in their severity and centrality estimates, (b) rater severity and centrality estimates were only moderately correlated (Study 1) or uncorrelated with each other (Study 2), (c) centrality effects had a demonstrable impact on examinee rank orderings, and (d) statistical indices of rater centrality derived from severity-only facets models (rater infit, residual-expected correlation, and standard deviation of rater-specific thresholds) correlated with centrality estimates much as predicted. The discussion focuses on implications for the analysis of rating quality in performance assessments.

Keywords: rater effects, rater centrality, facets models, rater-mediated assessment, Bayesian statistics, MCMC estimation

[1]*Correspondence concerning this article should be addressed to:* Thomas Eckes, PhD, TestDaF Institute, University of Bochum, Universitätsstr. 134, 44799 Bochum, Germany; e-mail: thomas.eckes@gast.de
[2]Hong Kong Examinations and Assessment Authority, Hong Kong

Human raters often judge the quality of examinee performances on assessment tasks using a rating scale. The scale categories describe different performance levels concerning a small set of criteria, aspects, or domains. The outcome of these kinds of assessments, variously called performance assessments (Kane et al., 1999; Lane & Iwatani, 2016) or rater-mediated assessments (Engelhard, 2002; McNamara, 2000), is a score or set of scores intended to represent the quality of the performance or the amount of the underlying proficiency.

The assessment of writing proficiency is a case in point. Tasks most commonly used elicit written performances (Eckes et al., 2016; Weigle, 2002, 2012). A prominent example is the timed impromptu essay: Examinees are given a topic (or prompt) and asked to write about it for a specified amount of time (e.g., 30, 45, or 60 minutes). Raters then assign a single score based on their overall impression of the performance (following a holistic approach); alternatively, they assign scores for distinct aspects of the performance separately (following an analytic approach). For example, raters assign separate scores for the aspects of content, organization, and language use.

It is a truism that raters, even if they are highly experienced, competent, and specially trained, are, to some extent, subject to various forms of errors and biases. The notorious subjectivity of human ratings threatens the validity of the use and interpretation of the scores assigned to examinees. Not surprisingly, there is a bulk of research addressing the question of how a satisfactory level of rating quality can still be achieved, including a wide array of statistical indices, quantitative methods, and conceptual approaches (Guilford, 1936; Gwet, 2014; Johnson et al., 2009; Saal et al., 1980; Wind & Peterson, 2018; Wolfe & McVay, 2012).

The term *rater effects* is used to summarize the different kinds of measurement error and bias that raters contribute to the outcomes of rater-mediated assessments (Myford & Wolfe, 2003; Wolfe & Song, 2016). In response to this topic's importance, rater effects and their implications for the design, analysis, and evaluation of rater-mediated assessment systems have been studied with increasingly sophisticated approaches (Eckes, 2015; Engelhard & Wind, 2018; Robitzsch & Steinfeld, 2018). This development is evidenced by recent special issues in *Psychological Test and Assessment Modeling* (Eckes, 2017, 2018), *Journal of Educational Measurement* (Engelhard & Wind, 2019), and *Applied Measurement in Education* (Wolfe & Wendler, 2020).

Here, we focus on a rater effect that has more and more attracted the attention of researchers in recent years: *central tendency* or *centrality*, and its opposite, *extreme response style* or *extremity* (Falk & Cai, 2016; Jin & Wang, 2014, 2018; Uto & Ueno, 2020; Wolfe & Song, 2015, 2016; Wu, 2017). Centrality occurs when raters provide scores that cluster around the midpoint of the rating scale; the opposite tendency, extremity, occurs when raters provide scores that are shifted towards the extreme categories or endpoints of the scale.

Psychometric and applied studies on centrality effects have mostly pursued one of two lines of research. The first line addresses the development, use, and interpretation of statistical indices of rater centrality based on a particular measurement model. The second

line aims to extend existing or develop new measurement models that incorporate a separate rater parameter explicitly representing centrality.

The present study addresses both lines of research. First, we briefly discuss several commonly-used centrality indices that rest on the many-facet Rasch measurement (MFRM) or facets modeling framework (Linacre, 1989). The focus then shifts to Jin and Wang's (2018) extension of facets models, including the estimation of a centrality parameter based on Bayesian statistics. Finally, we apply the Jin and Wang model to two data sets drawn from rater-mediated writing assessments and compare centrality parameter estimates to statistical indices of rater centrality.

Rater centrality indices under the facets modeling approach

Many statistical indices for detecting rater centrality in some way or other rest on the facets modeling approach. This approach has increasingly gained acceptance in a wide range of applied measurement contexts involving human raters (Aryadoust et al., 2021; Eckes, 2015, 2019; Engelhard & Wind, 2018; McNamara et al., 2019). Facets models are well suited for a principled, systematic analysis and evaluation of rater-mediated assessments at individual raters' level. Specifying raters as a separate facet of the assessment situation with other facets referring, for example, to examinees, scoring criteria, and tasks, allows researchers and practitioners to take a detailed look at each rater's susceptibility to various forms of errors and biases.

Facets model-severity only (FM-S)

One of the facets models in widespread use today is Linacre's (1989) many-facet extension of the Rasch rating scale model (RSM) proposed initially by Andrich (1978). Under this approach, rater severity effects are directly modeled by a severity parameter. Rater severity (or its opposite, leniency) is generally considered the most pervasive and detrimental effect; it manifests itself when raters provide scores that are consistently too low (or too high), compared to a group of raters or benchmark (criterion or expert) ratings. By contrast, rater centrality or other rater effects are not directly modeled; instead, they may be detected through some kind of post-hoc analysis, using statistical indices based on the output from a facets analysis. Adopting Jin and Wang's (2018) terminology, this class of MFRM models is called *facets model-severity only* or *FM-S*, for short.

In a three-facet assessment situation where J raters assign scores to N examinees on I criteria using a rating scale with $m + 1$ categories, that is, $k = 0, \ldots, m$, the following FM-S could be specified for studying rater effects:

$$\ln\left[\frac{p_{nijk}}{p_{nij(k-1)}}\right] = \theta_n - \beta_i - \alpha_j - \tau_k, \quad (1)$$

where p_{nijk} is the probability of examinee n receiving a rating of k from rater j on criterion i, $p_{nij(k\text{-}1)}$ is the probability of examinee n receiving a rating of k - 1 from rater j on criterion

i, θ_n is the ability of examinee n, β_i is the difficulty of criterion i, α_j is the severity of rater j, and τ_k is the difficulty of receiving a rating of k relative to k - 1.

In Equation 1, the parameter τ_k denotes the threshold parameter, which is defined as the location on the latent scale where the adjacent scale categories, k and k - 1, are equally probable to be observed. These locations are also called Rasch-Andrich thresholds (Andrich, 1998; Linacre, 2006). As indicated by the single subscript k of the threshold parameter, the model imposes the same rating scale structure on each rater and each criterion. That is, for all raters (and all criteria), the set of estimated threshold values will be the same. From a substantive point of view, this implies that across raters (and criteria), the categories of the rating scale are used in the same manner. There is some evidence supporting this assumption concerning the rating scale employed in the present research (Eckes, 2005, 2015).

Statistical indices of rater centrality

In the context of the FM-S (Eq. 1), *rater mean-square (MS) residual fit statistics*, or *rater fit statistics* for short, were among the first rater centrality indices discussed (Engelhard, 1992, 1994; Myford & Wolfe, 2003; Smith, 1996). Rater fit statistics indicate how well ratings provided by a given rater match the expected ratings generated by running a particular facets model. For example, residuals are computed as the difference (x_{nij} - e_{nij}) between the score that rater j assigns to examinee n on criterion i (the observed score x_{nij}) and the expected score (e_{nij}) based on the FM-S parameter estimates. The MS fit statistic for rater j is computed as the average of the squared standardized residuals over all examinees and criteria involved in producing that rater's scores (Eckes, 2015; Engelhard & Wind, 2018).

There are two different versions of rater fit statistics - rater infit and rater outfit. Rater infit, MS_W, is the *weighted* MS fit statistic; each squared standardized residual is weighted by its variance (or the amount of statistical information provided by the ratings). MS_W is sensitive to unexpected ratings where the locations of rater j and the other elements involved are aligned with each other, that is, close together on the measurement scale. Rater outfit, MS_U, is the *unweighted* MS fit statistic; it is sensitive to unexpected ratings where the latent variable locations of rater j and the locations of the other elements involved are farther apart from each other. Since unexpected ratings that provide more statistical information are generally associated with higher estimation precision, infit has been deemed more important than outfit for assessing rater fit (Linacre, 2002; Myford & Wolfe, 2003).

Rater fit statistics have an expected value of 1.0 and range from 0 to plus infinity (Linacre, 2002; Myford & Wolfe, 2003). Fit values greater than 1.0 indicate more variation than expected in the ratings. By contrast, fit values less than 1.0 indicate less variation than expected, meaning that the ratings tend to be muted, are too predictable, or provide redundant information; this is called *overfit*.

In operational settings, raters providing muted ratings in terms of overusing the middle category (or categories) of the rating scale have been associated with overfit. Such rating

tendencies would manifest themselves, for example, through MS_W values less than 0.75 (e.g., Engelhard, 1992, 1994). For this reason, rater overfit has been suggested as a potential indicator of rater centrality. However, more recent evidence has revealed that rater fit statistics may be sensitive to various other rater effects, in particular, halo effects, inaccuracy or randomness, and range of restriction; these statistics may also depend on properties of the observed score distributions (Myford & Wolfe, 2004; Wolfe et al., 2000, 2007). Therefore, any straightforward interpretation of rater fit statistics as rater centrality indices is called into question.

Wolfe (2004) suggested an alternative centrality index that similarly rests on residuals but follows different reasoning. When raters exhibit a centrality effect, the scores assigned to high-proficient examinees are lower than expected; hence, the residuals will be large and negative. Conversely, the scores assigned to low-proficient examinees are higher than expected; the residuals will be large and positive in this case. As a result, the Pearson correlation between the residual scores and the expected scores, that is, the residual-expected correlation, $r_{\text{res,exp}}$, will be negative: High expected scores tend to go with large negative residuals, and low expected scores tend to go with large positive residuals.

Rater fit statistics and residual-expected correlations rest on the FM-S rating scale model (Eq. 1). A third centrality index rests on the Rasch partial credit model (PCM) introduced by Masters (1982). In the present facets context, the partial credit FM-S version is given by

$$\ln\left[\frac{p_{nijk}}{p_{nij(k-1)}}\right] = \theta_n - \beta_i - \alpha_j - \tau_{jk}, \tag{2}$$

where all parameters are as in Equation 1 except for the τ_{jk} term. This term represents the difficulty of receiving a rating of k relative to k - 1 from rater j. Hence, Equation 2 specifies a rater-related three-facet partial credit model.

In Equation 2, the parameter τ_{jk} denotes the threshold parameter for a particular rater (indicated by the double subscript). In contrast to the FM-S (Eq. 1), it is no longer assumed that all raters share the same rating scale structure. Rather, each rater's rating scale is modeled to have its own category structure (for a discussion of rating scale vs. partial credit facets models, see Eckes, 2015, 2019; Linacre, 2000).

On this basis, a centrality index can be constructed as follows. When a particular rater exhibits a centrality effect, he or she tends to include a wide range of examinee proficiency levels in the middle category (or categories) of the rating scale. In this case, the lower thresholds will drop, and the higher thresholds will rise and, as a result, the average absolute distance between the rater-specific thresholds estimated using the FM-S (Eq. 2) will increase, relative to raters that are not subject to centrality.

Following this reasoning, the standard deviation of rater threshold parameter estimates, $SD(\tau_{jk})$, has been proposed as a centrality index (Myford & Wolfe, 2004; Wolfe & Song, 2015; Wu, 2017). Raters assigning scores associated with greater $SD(\tau_{jk})$ values are likely to exhibit a centrality effect; raters assigning scores associated with smaller $SD(\tau_{jk})$ values are likely to exhibit an extremity effect.

In a simulation study, Wolfe and Song (2015) compared rater fit statistics, residual-expected correlations, and rater threshold *SD* (or variance) in terms of their suitability for detecting rater centrality.[3] Results showed that the residual-expected correlation index demonstrated the best performance. More specifically, this index provided very low Type I error rates (i.e., incorrectly flagging raters who did not exhibit a centrality effect) and very low Type II error rates (i.e., not flagging raters who exhibited a centrality effect). These results were obtained under different levels of rater inaccuracy (randomness), centrality strength (magnitude of simulated centrality effects), and centrality prevalence (proportion of raters simulated to exhibit a centrality effect).

In a related study, Wolfe and Song (2014) considered two measurement models that differed according to whether or not they took local dependence between ratings into account. These models were (a) a standard rating scale facets model (Linacre, 1989) and (b) a random-effects facets model accounting for locally dependent ratings (Wang & Wilson, 2005). Local dependence between ratings may arise when multiple raters assign scores to the same examinee performance. Wolfe and Song found that residual-expected correlations remained highly consistent between these models. That is, classifying raters with $r_{res,exp} < -.30$ and $MS_W \leq 1.40$ as exhibiting a centrality effect, the flag rates were identical under the standard facets model and the random-effects facets model with perfect classification agreement between them.

Using rater threshold *SD* as a centrality indicator, Stafford et al. (2018) showed that this index performed very well even under large proportions of missing data, which are common in double-scoring rating designs. Finally, Song and Wolfe (2015) analyzed data containing multiple types of rater effects, that is, data that were simulated to represent the simultaneous occurrence of rater severity, centrality, and inaccuracy. Regarding the centrality detection, the rater threshold *SD* index was shown to produce low Type I and Type II error rates.

Modeling rater centrality: Broadening the facets perspective

The facets modeling approach discussed so far is limited in two respects. First, the only rater effect that is directly modeled is rater severity/leniency. Raters are arranged along the latent scale according to their severity estimates; similarly, examinees are arranged along that scale according to their proficiency estimates adjusted for the magnitude of between-rater severity differences. No other rater effect is taken into account. Second, the detection of rater effects over and above rater severity/leniency rests on statistical indices that are more of an indirect, post-hoc nature since they are derived after running a facets analysis. Without any doubt, these indices have a role to play in evaluating the psychometric quality of rater-mediated assessments. However, they do not impact the process of estimating examinee proficiencies.

[3]They also examined a fourth centrality index, rater slope (or discrimination), which is not discussed here because it is based on a two-parameter logistic model (an extension of the generalized partial credit model; Muraki, 1992) and, therefore, falls outside the class of MFRM models.

Myford and Wolfe (2004) have strongly emphasized the need to "refine existing computer programs (or develop new ones) that embody a more sophisticated approach to the detection of multiple rater effects and that will enable the adjustment of ratings for these multiple effects, not just for rater leniency/severity effects" (p. 220). The model advanced by Jin and Wang (2018) stands out by doing precisely that: accounting not only for rater severity but also for rater centrality.

Facets model-severity and centrality (FM-SC)

Jin and Wang's (2018) new facets model is an extended many-facet version of the partial credit model (Masters, 1982). For the present study, the following rating scale formulation of the original Jin and Wang model was used:

$$\ln\left[\frac{p_{nijk}}{p_{nij(k-1)}}\right] = \theta_n - \beta_i - \alpha_j - \omega_j\tau_k, \qquad (3)$$

where all parameters are as in Equation 1 except for the $\omega_j\tau_k$ term. In this term, the parameter ω_j (with $\omega_j > 0$) is a weight parameter representing the centrality of rater *j*. Hence, Equation 3 specifies a three-facet rating scale model accounting for both rater severity and centrality. Following Jin and Wang (2018), this class of facets model extensions is referred to as *facets model-severity and centrality* or *FM-SC*, for short.

The higher the values of ω_j, the more rater *j* tends to overuse the middle categories of the rating scale; conversely, the lower the values of ω_j, the more that rater tends to overuse the extreme categories of the rating scale. When $\omega_j = 1$ for $j = 1, \ldots, J$, the FM-SC reduces to the FM-S (Jin & Wang, 2018, p. 548).

FM-SC applications: Separating rater centrality from severity

In a simulation study, Jin and Wang (2018) demonstrated (a) the efficiency of the FM-SC to recover parameter estimates and (b) the consequences of ignoring rater centrality for parameter estimation. When raters exhibited unequal levels of centrality, the FM-SC parameter estimates were recovered very well. By contrast, fitting the FM-S to the simulation data, that is, ignoring rater centrality, yielded biased parameter estimates. When the more complex FM-SC was fitted to data where raters exhibited an equal level of centrality, that is, when the data were generated from the FM-S as the true model, parameter estimation remained mostly unaffected.

Besides conducting simulation studies, Jin and Wang (2018) illustrated FM-SC's practical utility with an English writing test taken by college students in Hong Kong. Similarly to Jin and Wang, the present research applied the FM-SC model (Eq. 3) to real data sets. The data sets were drawn from rater-mediated writing assessments administered in the context of admission to higher education institutions in Germany. In two independent studies, raters scored examinees' writing performances on a set of criteria using a four-category

rating scale. In addition to widening the scope of FM-SC applications, we aimed to investigate the relationship between centrality parameter estimates and indirect centrality indices under slightly different real assessment conditions.

Study 1 utilized a data set repeatedly examined adopting non-Bayesian (frequentist) approaches to fitting various instantiations of the FM-S (Eckes, 2015, 2019). In addition to possessing well-known characteristics, this data set has the advantage of being publicly available.[4] In Study 2, the rating data were collected as part of a validation program, focusing on the replicability of the outcomes from a high-stakes language assessment (for a related validation approach to the assessment of listening, see Eckes, 2020).

Differences between the two studies lay primarily in (a) the way raters were assigned to examinee performances and (b) the kind and number of criteria included in the scoring rubric (more detail on these differences is provided later).

Bayesian MCMC parameter estimation

The examinee, rater, and criterion parameters, as well as the Rasch-Andrich thresholds specified in the FM-S and FM-SC, respectively, were estimated building on a Bayesian approach (Gelman et al., 2013; Lunn et al., 2013). In particular, Bayesian parameter estimation was performed using Markov chain Monte Carlo (MCMC) techniques implemented in the JAGS freeware (JAGS = Just Another Gibbs Sampler; Plummer, 2017). The runjags package (Denwood, 2016, 2019) was employed to run the MCMC models in JAGS. This package provides interface functions to facilitate running user-specified MCMC models from within R (R Core Team, 2020). Also, runjags produces useful convergence diagnostics and a wide range of summary statistics right within the R environment.

Generally, Bayesian estimation methods involve modifying the likelihood function to incorporate any prior information known about model parameters, yielding a posterior distribution. From this distribution point estimates of parameters and the associated standard errors may be derived (for an introduction to Bayesian data analysis and psychometric modeling, see Jackman, 2009; Kruschke, 2015; Levy & Mislevy, 2016).

In a Bayesian approach, the model parameters are treated as random and assigned a prior distribution. Following Jin and Wang (2018), the prior distributions of all FM-SC parameters, except for the centrality parameter distribution, were assumed to be normal with a mean of 0 and a precision of 10 (precision is the inverse of the variance). More precisely, the prior distributions of the FM-SC parameters (see Eq. 3) were specified as follows:

[4]The complete data set is available at the following address: https://www.routledge.com/Quantitative-Data-Analysis-for-Language-Assessment-Volume-I-Fundamental/Aryadoust-Raquel/p/book/9781138733121#companion. The data are also available from the first author upon request.

$$\theta_n \sim N(0, 10), \quad (4)$$
$$\beta_i \sim N(0, 10), \quad (5)$$
$$\alpha_j \sim N(0, 10), \quad (6)$$
$$\tau_k \sim N(0, 10), \quad (7)$$
$$\omega_j \sim \text{lognormal}(0, 1), \quad (8)$$

where $N(\mu, \pi)$ designates the normal distribution with mean μ and precision π, for $\pi > 0$; the variance σ^2 of the normal distribution is $1/\pi$. Finally, lognormal(μ, π) designates the log-normal distribution, that is, the log transformation of a normal distribution with mean μ and precision π, for $\pi > 0$. The prior distributions of the four FM-S parameters (i.e., θ_n, β_i, α_j, and τ_k; see Eq. 1) were specified in the same way.

Three MCMC chains from different starting points were run to assess convergence to the posterior distribution. In each chain, the initial 5,000 draws were discarded as burn-in, and the draws from the subsequent 5,000 iterations were used for inference purposes, that is, retained for parameter estimation. The mean of the posterior distributions (based on a total of 15,000 draws) was used as the point estimate, or expected a-posteriori (EAP) estimate, of a given parameter; similarly, the posterior standard deviation was used as an estimate of the standard (or model) error associated with a parameter estimate. The gap between posterior draws was set at 10 to reduce the autocorrelation effect; that is, every 11th posterior draw was recorded. This set of specifications was the same for estimating the FM-SC and the FM-S parameters, respectively.

As an index of convergence to the posterior distribution, the proportional scale reduction factor (PSRF) of the Gelman-Rubin statistic (Gelman & Rubin, 1992) was computed (included in the runjags package as a default option). The PSRF index compares, for each parameter, the between-chain and within-chain variances of samples from the posterior distribution. It is commonly suggested to infer that the chains have converged to the posterior distribution if the PSRF value is close to 1 (i.e., PSRF < 1.1; Levy & Mislevy, 2016, p. 109).

The posterior predictive model-checking (PPMC) method (Rubin, 1984) was used to examine the fit of the (observed) data to the model. This method compares the observed data with the data that are generated or predicted by the model (Gelman et al. 2013; Sinharay, 2005). In particular, the PPMC approach involves computing a discrepancy measure using each simulated value from the posterior distributions for the parameters. Plotting the distribution of these values (the realized values) against the posterior predictive values' distribution provides a graphical display of data-model fit, which may be summarized in the tail-area probability, also known as the posterior predictive *p*-value (PPP-value). Extreme PPP-values (i.e., values close to 0 or 1) can be considered to indicate poor data-model fit; medium values, that is, values around .5, indicate much better fit (Levy & Mislevy, 2016, p. 242).

Finally, to address the issue of relative model fit, that is, to compare the FM-SC to the FM-S in terms of data-model fit, the Bayesian deviance information criterion (DIC; Spiegelhalter et al., 2002) was computed for each model:

$$DIC = \check{D} + p_D \,, \tag{9}$$
$$pD = \check{D} - D^*, \tag{10}$$

where $\check{D}$ is the posterior mean of the deviance, p_D is the penalty for model complexity, and D^* is the deviance evaluated at the posterior mean. Models showing smaller DIC values are preferred as better fitting (Levy & Mislevy, 2016, p. 248). The DIC statistic is available in the runjags package using the extract.runjags function.

Research questions

In this research, two writing assessment studies served to illustrate a Bayesian approach to the estimation of rater severity and centrality parameters. The studies shared the following methodological key elements: (a) FM-SC analyses were run to simultaneously measure rater severity and centrality using a Bayesian MCMC approach; (b) FM-S analyses were run on the same data using the same approach for purposes of comparison with the FM-SC findings; (c) non-Bayesian facets analyses were run on the same data building on the FM-S; (d) statistical indices of rater centrality were computed based on the non-Bayesian FM-S analyses and compared to the Bayesian FM-SC centrality measures.

Building on this methodological approach, the present research aimed to answer the following three research questions:

1. Do raters differ in their severity/leniency and centrality/extremity when scoring examinee writing performances, and, if so, how pronounced are these differences?
2. What is the impact of between-rater centrality differences on examinee writing proficiency estimates? Put differently, how much do examinee rank orders derived from the FM-SC writing proficiency estimates agree with those derived from the FM-S estimates?
3. How do estimates of the ω-parameter compare to statistical indices of centrality effects? More specifically, what is the relationship between Bayesian FM-SC ω-estimates and (non-Bayesian) rater infit statistics, residual-expected rating correlations, and rater threshold *SD*? Specifically, in the light of the above discussion, the expectation is that the ω-estimates will correlate (a) significantly negatively with the residual-expected rating correlations and (b) significantly positively with rater threshold *SD*. Regarding the infit (*MSw*) statistic, the situation is less clear. However, given the current operational setting, it appears reasonable to expect negative correlations with ω-estimates.

Method

Participants

In both studies, the examinees were international students applying for entry to higher education institutions in Germany. Raters were specialists in German as a foreign language who had been trained and monitored as to compliance with the scoring guidelines. A total of 307 examinees (149 males, 158 females) completed the Study 1 writing task; the writing performances were evaluated by a group of 18 raters (4 males, 14 females). The Study 2 writing task was completed by 206 examinees (66 males, 140 females); ratings were provided by a group of 12 raters (1 male, 11 females).

Instruments and procedure

In each study, the writing task was part of the Test of German as a Foreign Language (TestDaF, *Test Deutsch als Fremdsprache*). The TestDaF is officially recognized as a language exam for international students applying for entry to higher education institutions in Germany (Eckes & Althaus, 2020). Examinee performance in each of four test sections (reading, listening, writing, and speaking) is related to one of three levels of language proficiency, the so-called TestDaF levels (*TestDaF-Niveaus*, TDNs). The levels TDN 3, TDN 4, and TDN 5 cover the Council of Europe's (2001) Lower Vantage Level (B2.1) to Higher Effective Operational Proficiency (C1.2); that is, the test measures German language proficiency at an intermediate to a high level. Examinees achieving at least TDN 4 in each section are eligible for admission to a German institution of higher education (for a review of the TestDaF, see Norris & Drackert, 2018; see also https://www.testdaf.de, where free sample tests are available).

The TestDaF writing section (duration: 60 min) assesses an examinee's ability to produce a coherent and well-structured text on a given topic taken from the academic context. There is a single task, requiring two types of prose: description and argumentation. More precisely, in the first part of this section, charts, tables, or diagrams are provided along with a short introductory text, and the examinee is asked to describe the relevant information. Specific points to be dealt with are stated in the rubric. In the second part, the examinee has to consider different positions on an aspect of the topic and write a well-structured argument. The input consists of short statements, questions, or quotes. As before, aspects to be dealt with in the argumentation are stated in the rubric.

The rating design used in Study 1 was as follows. Each performance on the writing task, that is, each essay, was rated independently by two raters. Also, one rater provided ratings of two essays that were randomly selected from each of the other 17 raters' workload. This design served to satisfy the essential requirement of a connected dataset in which all elements of all facets, that is, examinees, raters, and scoring criteria, are directly or indirectly linked to each other (Eckes, 2015, 2019).

Raters scored the essays on the following three criteria: *global impression* (referring to lower-level aspects such as fluency, train of thought, and structure), *task fulfillment* (completeness, description, and argumentation), and *linguistic realization* (breadth of syntactic elements, vocabulary, and correctness). For each criterion, raters were provided with scale descriptors specifically designed to characterize written performance at each rating scale category, that is, at each TDN.

For example, the descriptors for the global impression aspect of fluency were as follows: "On the whole, the text does not read fluently" (*below TDN 3*), "Repeated reading of parts of the text is necessary" (*TDN 3*), "Readability is slightly impaired in places" (*TDN 4*), and, finally, "The text reads fluently throughout" (*TDN 5*). The descriptor-specific TDNs were aggregated to yield a single TDN score for each of the three criteria. For computations, *below TDN 3* was scored "2", and the other levels were scored from "3" to "5".

On each criterion, there were 648 ratings; that is, 614 double ratings plus 34 third ratings, making a total of 1,944 ratings (the proportion of missing ratings was 88.3%). These ratings provided the input for estimating parameters for raters, examinees, criteria, and the Rasch-Andrich thresholds based on the FM-SC and FM-S.

The Study 2 rating data were collected as part of an ongoing validation program, focusing on the TestDaF writing, speaking, and listening sections. The writing performances were sampled from the entire set of 3,949 essays produced by TestDaF examinees in April 2012 (2,557 females, 1,392 males). For each examinee, two kinds of data were available: (a) a TDN rating on each criterion, and (b) a final TDN level for writing. The range of examinee proficiencies most critical in terms of eligibility for university admission (i.e., TDN levels 3 and 4) was covered by randomly drawing 100 examinees from among examinees who scored near the borderline for each of these levels at the lower end of the TDN scale (i.e., below TDN 3 vs. TDN 3 and TDN 3 vs. TDN 4); another 100 examinees were drawn from the entire group completely at random. Accidentally, six further examinees were randomly selected and included in the present sample of examinees.

Similar to Study 1, the rating design used in Study 2 was incomplete but connected. Differences from Study 1 lay in the kind of assignment of raters to performances. In particular, all 12 raters independently rated the same subset of 10 randomly selected essays; most of the remaining essays were each rated by a single rater (some of these essays were also rated by two raters each to strengthen the link between raters).

Another significant difference from Study 1 concerned the set of criteria used for scoring examinee performances. Raters in Study 2 scored the essays on each of the lower-level aspects *fluency*, *train of thought*, and *structure* (replacing the higher-level *global impression* criterion); *completeness*, *description*, and *argumentation* (replacing the *task fulfillment* criterion); and *breadth of syntactic elements*, *vocabulary*, and *correctness* (replacing the *linguistic realization* criterion). For each of these nine criteria, raters were provided with the same scale descriptors used in Study 1.

One rater inadvertently returned scores for only 29 examinees; the other 11 raters provided scores for 30 examinees, resulting in a set of 359 ratings on each of the nine criteria. Thus, a total of 3,231 ratings was available for estimating FM-SC and FM-S parameters (the proportion of missing ratings was 85.5%).

Data analysis

In both studies, the performance ratings on the four-category rating scale (rescored from 1 to 4) provided the input to the Bayesian MCMC estimation of the FM-SC (and the FM-S) parameters using the runjags package (Denwood, 2016). Unlike Jin and Wang's (2018) sample data, the raters were all operational raters; that is, there was no expert rater who could have been treated as a reference or an anchor. Hence, for model identification, the severity distribution's mean was set to 0, and the centrality distribution's mean was set to 1. Following the same rationale, before running the Bayesian FM-S parameter estimation the severity distribution's mean was set to 0.

Regarding the non-Bayesian approach to the three-facet analysis, the computer program FACETS (Version 3.83; Linacre, 2020) was used. For more than three decades, FACETS has enjoyed great popularity in the field of language assessment and beyond (Aryadoust et al., 2021; Eckes, 2015, 2019; Engelhard & Wind, 2018; McNamara et al., 2019). This program can accommodate applications of a wide range of Rasch models for rating data, including the severity facets model (FM-S; Eq. 1); however, it does not allow for estimating rater centrality. FACETS uses joint maximum likelihood (JML) estimation of the model parameters.

Specifying the rating scale variant of the FM-S in FACETS provided, for each rater, the first statistical index of rater centrality, that is, the infit mean-square statistic (MS_W). Also, from the output of this analysis, the correlation between the expected and residual scores ($r_{exp,res}$) was computed for each rater. Finally, a partial credit model variant was specified, where the rating scale for each rater was modeled to have its own category structure (partial credit FM-S; Eq. 2). From the FACETS output of this analysis, the standard deviation of the rater-related Rasch-Andrich thresholds was computed to yield the third statistical index of rater centrality (SD_τ).

Results

Data-model fit

Convergence and data-model fit statistics are summarized in Table 1. Across studies, for each parameter under the FM-S and the FM-SC, respectively, the potential scale reduction factor (PSRF) values were very close to 1.0, indicating that the three MCMC chains converged to the target (posterior) distribution without problems. Also, the PPP-values were much greater than 0, confirming that, in each instance, the data-model fit was satisfactorily high. Finally, the DIC statistic values revealed that the FM-SC fit the data better than the FM-S, taking into account the greater number of estimated parameters in terms of the penalty statistic for the FM-SC.

Table 1:
Bayesian data-model fit statistics for the FM-S and the FM-SC in Study 1 and Study 2 writing assessments

Statistic	FM-S	FM-SC
	Study 1	
PSRF (min-max)		
Examinee proficiency	1.000-1.003	1.000-1.005
Rater severity	1.000-1.007	1.001-1.023
Rater centrality	-	1.000-1.013
Criterion difficulty	1.002	1.005-1.007
Rasch-Andrich Thresholds	1.001-1.002	1.001-1.003
PPP-value	0.376	0.344
DIC	3,202.4	3,195.1
Deviance ($\check{D}$)	2,902.6	2,880.9
Penalty (p_D)	299.8	314.2
	Study 2	
PSRF (min-max)		
Examinee proficiency	1.000-1.003	1.000-1.003
Rater severity	1.000-1.005	1.001-1.005
Rater centrality	-	1.000-1.003
Criterion difficulty	1.001-1.002	1.001
Rasch-Andrich Thresholds	1.000-1.002	1.000-1.003
PPP-value	0.262	0.280
DIC	5,846.7	5,814,2
Deviance ($\check{D}$)	5,636.8	5,593.9
Penalty (p_D)	209.9	220.3

Note. Throughout the FM-S and FM-SC analyses, the rating scale versions were used. PSRF = Proportional scale reduction factor. PPP-value = Posterior predictive *p*-value. DIC = Deviance information criterion.

Bayesian rater parameter estimates

Tables 2 and 3 present the FM-S and the FM-SC rater parameter estimates. Raters are ordered in the tables by centrality measures (ω estimates), from high to low. Also shown are the summary statistics for the observed scores based on the TDN rating scale (original ratings from 2 to 5).

Regarding Study 1 (Table 2), the FM-S rater severity measures (α estimates) had a 4.46-logit spread; Rater 16 was the most severe rater, and Rater 1 the most lenient rater. Not surprisingly, the observed averages correlated highly significantly with the FM-S rater severity measures, $r(18) = -.92$, $p < .001$. Concerning the FM-SC, the rater severity measures' spread was 5.27 logits; again, the severity measures were highly significantly correlated with the observed averages, $r(18) = -.93$, $p < .001$. The correlation between FM-S and FM-SC severity measures was close to 1, $r(18) = .99$, $p < .001$, attesting to the high stability of severity estimates across models.

As to the ω-estimates, it can be seen that the rater centrality measures ranged from 1.44 (Rater 1) to 0.69 (Rater 16). Between-rater centrality differences of this magnitude are

likely to impact on the final scores awarded to examinees (as will be shown later). In line with expectations, there was a strong negative correlation between the centrality estimates and the standard deviations of the observed scores, $r(18) = -.91$, $p < .001$; that is, the smaller the dispersion of a given rater's observed score distribution, the greater the centrality estimate of that rater (Jin & Wang, 2018, reported a highly similar finding). Appendix 1 presents the observed score distributions for the complete set of 18 raters with the associated centrality estimates.

Table 2:
Bayesian measurement results for 18 raters in the Study 1 writing assessment using the FM-S and the FM-SC

	Observed scores			FM-S (RSM)	FM-SC (RSM)	
Rater	*N*	*M*	*SD*	α Est. (*SE*)	α Est. (*SE*)	ω Est. (*SE*)
1	60	4.52	0.54	-2.17 (.41)	-3.08 (.85)	1.44 (.30)
17	135	3.98	0.77	-0.62 (.28)	-0.74 (.32)	1.28 (.12)
4	72	3.72	0.76	0.04 (.42)	0.03 (.42)	1.27 (.15)
9	141	3.39	0.83	1.02 (.26)	1.20 (.27)	1.24 (.11)
7	204	4.06	0.77	-1.93 (.25)	-2.01 (.28)	1.22 (.10)
3	123	4.02	0.82	-1.37 (.29)	-1.36 (.32)	1.11 (.12)
14	129	3.45	0.84	1.46 (.29)	1.56 (.31)	1.10 (.11)
6	102	3.59	0.93	0.10 (.22)	0.19 (.22)	1.10 (.10)
11	57	3.54	0.95	0.05 (.38)	0.07 (.41)	1.03 (.14)
13	123	3.11	0.98	2.03 (.32)	2.08 (.32)	1.00 (.11)
8	141	3.49	0.99	0.28 (.26)	0.40 (.27)	0.97 (.09)
12	132	3.61	0.94	-0.48 (.27)	-0.34 (.26)	0.96 (.10)
10	123	3.48	0.97	-0.68 (.27)	-0.64 (.27)	0.94 (.10)
2	72	4.10	0.86	-1.28 (.42)	-1.00 (.44)	0.87 (.13)
18	63	3.81	1.01	-0.31 (.38)	-0.20 (.37)	0.82 (.13)
15	84	3.58	1.00	0.79 (.33)	0.91 (.33)	0.76 (.11)
5	123	3.37	1.06	0.79 (.29)	0.73 (.28)	0.73 (.09)
16	60	3.03	1.10	2.29 (.41)	2.19 (.42)	0.69 (.14)

Note. Observed scores refer to the four-category rating scale ranging from 2 (*below TDN 3*) to 5 (*TDN 5*). FM-S and FM-SC rater estimates were computed according to the many-facet rating scale model using a Bayesian approach. *N* is the number of ratings. α Est. is the estimate of the rater severity parameter. ω Est. is the estimate of the rater centrality parameter. Raters are ordered by centrality estimates, from high (centrality) to low (extremity).

Importantly, rater severity and centrality estimates were only moderately correlated with each other, $r(18) = -.52$, $p < .05$. Therefore, rater centrality estimates appear to provide information about scoring tendencies not captured by rater severity measures. For example, Raters 4 and 18 exhibited different levels of centrality (1.27 vs. 0.82), but similar levels of severity, both of which were around the origin of the logit scale (0.03 and -0.20, respectively); conversely, Raters 3 and 14 exhibited similar levels of centrality (1.11 and 1.10, respectively) but widely different levels of severity (-1.36 vs. 1.56).

Table 3:
Bayesian measurement results for 12 raters in the Study 2 writing assessment using the FM-S and the FM-SC

	Observed scores			FM-S (RSM)	FM-SC (RSM)	
Rater	*N*	*M*	*SD*	α Est. (*SE*)	α Est. (*SE*)	ω Est. (*SE*)
1	270	3.39	0.77	0.64 (.15)	0.72 (.16)	1.32 (.08)
12	270	3.34	0.88	-0.75 (.16)	-0.76 (.16)	1.20 (.08)
10	270	3.49	0.87	0.06 (.14)	0.07 (.15)	1.15 (.07)
5	260	3.51	0.83	0.25 (.15)	0.24 (.15)	1.13 (.08)
3	270	3.41	0.94	0.12 (.11)	0.12 (.11)	1.04 (.06)
6	270	3.83	0.88	-1.54 (.16)	-1.54 (.17)	1.03 (.08)
7	270	3.53	0.93	-0.31 (.15)	-0.30 (.15)	1.02 (.07)
8	270	3.36	0.95	0.07 (.14)	0.06 (.14)	0.88 (.07)
2	270	3.36	1.00	0.28 (.11)	0.26 (.11)	0.88 (.06)
4	270	3.39	0.94	-0.11 (.15)	-0.10 (.15)	0.87 (.07)
9	270	3.01	0.91	0.99 (.15)	0.93 (.15)	0.86 (.07)
11	270	3.46	1.01	0.30 (.15)	0.30 (.15)	0.80 (.06)

Note. Observed scores refer to the four-category rating scale ranging from 2 (*below TDN 3*) to 5 (*TDN 5*). FM-S and FM-SC rater estimates were computed according to the many-facet rating scale model using a Bayesian approach. *N* is the number of ratings. α Est. is the estimate of the rater severity parameter. ω Est. is the estimate of the rater centrality parameter. Raters are ordered by centrality estimates, from high (centrality) to low (extremity).

Figure 1 illustrates how rater severity and centrality jointly impact the observed score distributions (Study 1 data). The figure also demonstrates that both parameter estimates must be taken into account when drawing conclusions concerning a given rater's influence on the observed ratings. Compare again Rater 4 (average severity, high centrality) and Rater 18 (about average severity, low centrality). Since both raters are similarly severe, it seems safe to conclude that Rater 4 is more subject to central tendencies than Rater 18. As another example, consider Rater 3 (low severity, high centrality) and Rater 14 (high severity, high centrality). Given the close correspondence in these raters' central tendencies, the clearly different severity estimates provide evidence that Rater 3 tends to assign much more lenient ratings than Rater 14. Finally, there is considerably less confidence in interpreting raters' influence on the scores they assign to examinees when raters show both different levels of severity and different levels of centrality, as with Rater 2 (low severity, low centrality) and Rater 9 (high severity, high centrality).

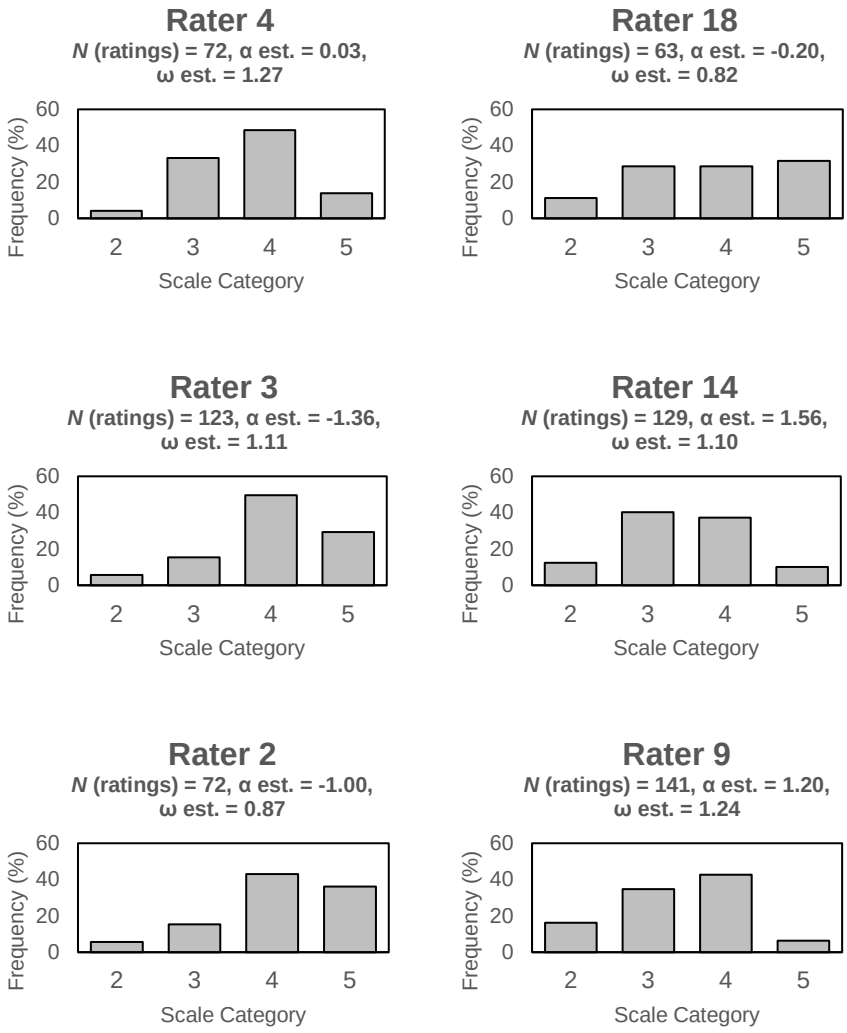

Figure 1:
Illustrative frequency plots for six raters (Study 1).

In Study 2 (Table 3), the FM-S rater severity measures had a 2.53-logit spread. The observed averages correlated highly significantly with the FM-S rater severity measures, $r(12) = -.76$, $p < .01$. Concerning the FM-SC, the rater severity measures' spread was 2.47 logits; again, the severity measures were highly significantly correlated with the observed

averages, $r(12) = -.74$, $p < .01$. Confirming the Study 1 finding, the correlation between FM-S and FM-SC severity measures was close to 1, $r(12) = .999$, $p < .001$.

As can be seen from the ω estimates, central tendencies were less dispersed than among Study 1 raters, ranging from 1.32 (Rater 1) to 0.80 (Rater 11). There was a strong negative correlation between the centrality estimates and the standard deviations of the observed scores, $r(12) = -.88$, $p < .001$; that is, the smaller the observed standard deviation, the greater the estimated rater centrality. Appendix 2 presents the observed score distributions for the complete set of 12 raters.

The correlation between rater severity and rater centrality estimates was not statistically significant, $r(12) = -.11$, *ns*. Thus, even more so than in Study 1, rater centrality diverged from rater severity. For example, Raters 10 and 8 showed different levels of centrality (1.15 vs. 0.88), but similar levels of severity, both of which were around the origin of the logit scale (0.07 and 0.06, respectively); conversely, Raters 3 and 6 showed similar levels of centrality (1.04 and 1.03, respectively) but widely different levels of severity (0.12 vs. -1.54).

Impact of rater centrality on examinee proficiency estimates

The examinee proficiency measures estimated under the two models were highly correlated with each other (close to 1), $r(307) = .997$, $p < .001$ (Study 1), $r(206) = .996$, $p < .001$ (Study 2). Figure 2 displays the corresponding scatter diagrams, with FM-SC estimates shown on the horizontal axis and FM-S estimates on the vertical axis. Though, at first sight, this finding may seem to suggest that the impact of rater centrality on estimated examinee proficiency is negligibly small, the agreement between FM-SC and FM-S estimates appears to be somewhat less pronounced among examinees with very high or very low writing proficiency.

Following the approach taken by Jin and Wang (2018), the Study 1 examinee rank-ordering resulting from the FM-SC proficiency estimates (as a reference) was compared to the Study 1 examinee rank-ordering produced by the FM-S estimates. This comparison yielded an absolute rank-order difference ranging from 0 to 24 ($M = 4.01$, $SD = 3.92$). On average, the rank orderings of examinees differed by four ranks, depending on which model was used for estimating their proficiency. Rank differences of this magnitude may have serious consequences for individual examinees, for example, in selection decisions. Thus, if the top 20 out of 307 examinees were admitted to a prestigious course or field of study, then three of the most qualified examinees would lose out on this opportunity when their proficiency was estimated based on the FM-S instead of the FM-SC; in case of a less strict selection criterion, say admitting the top 50 examinees, three examinees would still be put at a disadvantage. In Study 2, the corresponding absolute rank-order difference ranged from 0 to 15 ($M = 3.55$, $SD = 2.94$). On average, the examinee rank orderings differed by three-and-a-half ranks, depending on whether the FM-SC or the FM-S was used for estimating examinee proficiency.

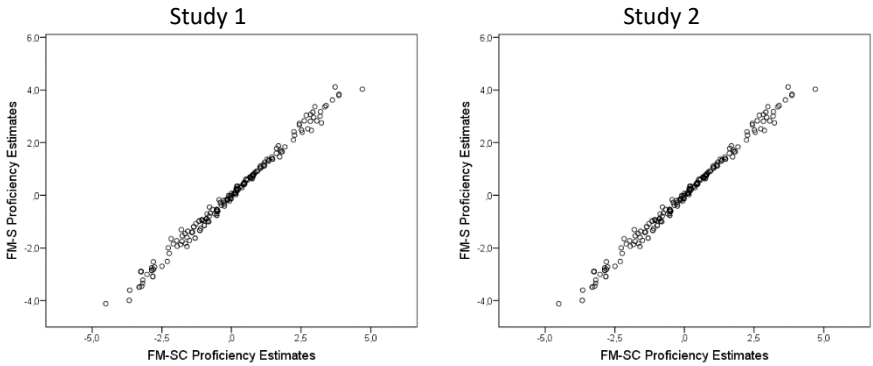

Figure 2:
Relationship between FM-SC and FM-S examinee proficiency estimates.

Table 4:
Bayesian FM-SC rater centrality estimates compared to non-Bayesian FM-S centrality indices (Study 1)

	FM-SC (RSM)	FM-S (RSM)		FM-S (PCM)
Rater	ω Est. (*SE*)	MS_W	$r_{res,exp}$	SD_τ
1	1.44 (.30)	0.96	-.09	3.48
17	1.28 (.12)	0.81	-.05	4.52
4	1.27 (.15)	0.89	-.14	4.20
9	1.24 (.11)	0.81	-.04	4.46
7	1.22 (.10)	0.94	-.12	4.61
3	1.11 (.12)	0.82	.03	3.75
14	1.10 (.11)	1.10	-.04	3.81
6	1.10 (.10)	1.11	-.07	4.02
11	1.03 (.14)	0.75	.20	3.78
13	1.00 (.11)	0.82	-.03	4.35
8	0.97 (.09)	1.05	.01	3.67
12	0.96 (.10)	1.08	.04	3.57
10	0.94 (.10)	1.02	.12	3.50
2	0.87 (.13)	1.16	.16	2.78
18	0.82 (.13)	1.30	-.07	3.20
15	0.76 (.11)	1.39	.09	2.67
5	0.73 (.09)	1.12	.14	2.67
16	0.69 (.14)	0.93	.17	3.35

Note. ω Est. is the estimate of the rater centrality parameter computed according to the FM-SC rating scale model using a Bayesian approach. MS_W is an information-weighted mean-square fit statistic (infit) computed according to the FM-S rating scale model using FACETS. $r_{res,exp}$ is the Pearson correlation between expected scores and residuals computed according to the FM-S rating scale model using FACETS. SD_τ is the standard deviation of the Rasch-Andrich thresholds computed according to the FM-S partial credit model using FACETS. Raters are ordered by centrality estimates, from high (centrality) to low (extremity).

Bayesian centrality estimates vs. non-Bayesian centrality indices

Tables 4 and 5 display the non-Bayesian rater centrality indices provided by the FACETS program (MS_W) or computed from the FACETS output ($r_{res,exp}$, SD_τ) in Study 1 and Study 2, respectively. For ease of comparison, the ω estimates from the Bayesian FM-SC analysis are also shown. Again, raters are ordered by their ω estimate (from high to low). Table 6 presents the correlations between these different approaches to assess rater centrality.

Across studies, a fairly consistent pattern of correlations emerged. In particular, the correlations between the ω estimate and the residual-expected correlation and the SD_τ index, respectively, were statistically highly significant and, as expected, in opposite directions. Thus, higher central tendencies estimated under the FM-SC were associated with lower (typically negative) correlations between expected and residual scores on the one hand, and with higher standard deviations of the Rasch-Andrich threshold estimates on the other.

There were also some weaker correlations. These correlations were primarily associated with the infit statistic. In Study 1, infit was only moderately correlated with the ω estimate and uncorrelated with the residual-expected correlation index; in Study 2, infit was not statistically significantly correlated with any other centrality estimates or indices.

Table 5:
Bayesian FM-SC rater centrality estimates compared to non-Bayesian FM-S centrality indices (Study 2)

	FM-SC	FM-S (RSM)		FM-S (PCM)
Rater	ω Est. (*SE*)	MS_W	$r_{res,exp}$	SD_τ
1	1.32 (.08)	0.98	-.18	3.56
12	1.20 (.08)	0.85	-.09	3.39
10	1.15 (.07)	0.88	-.06	3.08
5	1.13 (.08)	1.00	-.08	2.93
3	1.04 (.06)	0.89	.01	2.76
6	1.03 (.08)	0.93	.02	2.69
7	1.02 (.07)	1.16	-.07	2.74
8	0.88 (.07)	0.93	.12	2.36
2	0.88 (.06)	0.80	.21	2.35
4	0.87 (.07)	0.98	.13	2.25
9	0.86 (.07)	1.12	.05	2.24
11	0.80 (.06)	1.46	-.04	2.15

Note. ω Est. is the estimate of the rater centrality parameter computed according to the FM-SC rating scale model using a Bayesian approach. MS_W is an information-weighted mean-square fit statistic (infit) computed according to the FM-S rating scale model using FACETS. $r_{res,exp}$ is the Pearson correlation between expected scores and residuals computed according to the FM-S rating scale model using FACETS. SD_τ is the standard deviation of the Rasch-Andrich thresholds computed according to the FM-S partial credit model using FACETS. Raters are ordered by centrality estimates, from high (centrality) to low (extremity).

Table 6:
Pearson correlations between Bayesian FM-SC rater centrality estimates and non-Bayesian FM-S centrality indices

	ω Est.	MS_W	$r_{res,exp}$
	Study 1		
MS_W	-.55*		
$r_{res,exp}$	-.71**	.10	
SD_τ	.72**	-.72**	-.63**
	Study 2		
MS_W	-.41		
$r_{res,exp}$	-.77**	-.25	
SD_τ	.99**	-.41	-.77**

Note. Estimates of the centrality parameter ω were computed according to the FM-SC rating scale model using a Bayesian approach. MS_W is an information-weighted mean-square fit statistic (infit) computed according to the FM-S rating scale model using FACETS. $r_{res,exp}$ is the Pearson correlation between expected scores and residuals computed according to the FM-S rating scale model using FACETS. SD_τ is the standard deviation of the Rasch-Andrich thresholds computed according to the FM-S partial credit model using FACETS. * $p < .05$. ** $p < .01$.

Bayesian estimates of criterion difficulty and Rasch-Andrich thresholds

Though not the focus of the present research, the criterion difficulty and Rasch-Andrich threshold parameters yielded further evidence on the efficiency of the basic MCMC estimation procedure. Table 7 presents Study 1 and Study 2 criterion parameter estimates based on the FM-SC. Remember that in Study 2, the three higher-level criteria used in Study 1 were each subdivided into three lower-level aspects. In Study 1, *global impression* was by far the easiest criterion, followed by *task fulfillment* and *linguistic realization*, which did not differ significantly from one another in estimated difficulty. In Study 2, *structure* and *completeness* were the easiest lower-level aspects; *train of thought*, *correctness*, and *description* were the most difficult ones.

Table 7:
Bayesian FM-SC (RSM) estimates of criterion difficulty

Criterion	β Est. (*SE*)
Study 1	
Global impression	-1.38 (.18)
Task fulfillment	-0.19 (.18)
Linguistic realization	-0.11 (.18)
Study 2	
Fluency	0.41 (.16)
Train of thought	0.73 (.16)
Structure	-0.19 (.16)
Completeness	-0.13 (.16)
Description	1.17 (.16)
Argumentation	0.69 (.16)
Breadth of syntactic elements	0.19 (.16)
Vocabulary	0.40 (.16)
Correctness	0.88 (.16)

The FM-SC threshold parameter estimates (Table 8) confirm that the four rating scale categories functioned as intended. According to Linacre (2004), Rasch-Andrich thresholds should advance monotonically with categories by at least 1.4 logits and, at the same time, by less than 5.0 logits (see also Eckes, 2015). As shown in Table 8, in both Study 1 and Study 2, the differences between threshold values for adjacent categories all stayed within the range defined by these lower and upper control limits.

Table 8:
Bayesian FM-SC (RSM) estimates of Rasch-Andrich thresholds

Threshold	Study 1	Study 2
τ_1	-2.95 (.15)	-2.44 (.08)
τ_2	-0.07 (.08)	-0.07 (.06)
τ_3	3.02 (.14)	2.51 (.09)

Note. Values in parentheses denote standard errors (*SE*) of the threshold estimates.

Summary and discussion

This research adopted a Bayesian MCMC approach to investigating rating quality in the context of rater-mediated writing assessments. Specifically, the facets model-severity and centrality (FM-SC) proposed by Jin and Wang (2018) was used to estimate individual raters' severity/leniency and centrality/extremity tendencies within the same many-facet Rasch measurement framework. The data came from two different three-facet assessment situations with independent samples of examinees and raters and different sets of scoring criteria (Study 1, Study 2). In each study, raters scored examinee writing performances on a four-category rating scale.

The FM-SC analysis was conducted from within the R environment (R Core Team, 2020) using the R package runjags (Denwood, 2016, 2019). Three chains were run to estimate

model parameters and to provide convergence diagnostics. As evidenced by the values of the proportional scale reduction factor (PSRF) computed for each model parameter, the chains converged to the posterior distribution without any problem. Adopting the same Bayesian measurement framework, the less complex severity facets model (FM-S), which does not include a centrality parameter, was applied to the essay rating data for comparison purposes. The posterior predictive *p*-values (PPP-values) indicated that both models (FM-SC and FM-S) had satisfactorily high data-model fit. Concerning their relative fit, the deviance information criterion (DIC) provided evidence that the FM-SC had an advantage over the FM-S.

In both studies, raters were clearly separated along the centrality dimension. Whereas some raters showed a marked tendency to assign scores around the scale midpoint (TDN 3 or TDN 4), indicated by an ω-estimate much greater than 1, other raters tended toward the extreme ends of the rating scale (below TDN 3 vs. TDN 5), indicated by an ω-estimate much smaller than 1. There was only a moderate correlation (Study 1) and a non-significant correlation (Study 2), respectively, between rater centrality and severity measures. This finding supports the view that both rater effects should be simultaneously measured when analyzing rater-mediated assessments.

Even though examinee proficiency measures estimated under the FM-SC and the FM-S were very highly correlated with each other, the resulting rank orders of examinees sorted from high to low proficiency differed on average by four (Study 1) or three-and-a-half ranks (Study 2). This may prove to be a critical difference for some examinees regarding course admissions or other high-stakes decisions. In any case, the shifts in examinee rankings demonstrate that even small differences in model-specific proficiency estimates can have practically relevant implications. In line with this finding, Wind (2019) showed that rater centrality effects substantially influence estimates of examinee proficiency and classification decisions (though to a somewhat lesser degree than rater severity and inaccuracy effects).

Within the framework of many-facet Rasch measurement, the usual way to deal with centrality effects is to run a rating scale or partial credit analysis and use some statistical index as a post-hoc method to gauge the extent to which raters exhibited centrality or extremity in their ratings. The correlations between the ω-estimate and three popular statistical indices (mean-squares infit, residual-expected correlation, and the threshold standard deviation) provided evidence that some indices are much more suitable for this purpose than others. In particular, the mean-squares infit statistic (MS_W) was only moderately correlated with the ω-estimate; the correlation with the residual-expected correlation ($r_{\text{res,exp}}$) was small and non-significant.

By comparison, both the $r_{\text{res,exp}}$ and the standard deviation (SD_τ) indices correlated strongly and significantly with the ω-estimate. This finding confirms more recent concerns against the indiscriminate use and interpretation of mean-squares statistics as indicators of rater centrality (Myford & Wolfe, 2004; Wolfe & Song, 2015, 2016).

More importantly, as discussed above, no matter which post-hoc statistic may be computed in a given research context, none provides a direct measure of centrality effects following the same psychometric reasoning as the modeling of rater severity effects. In contrast,

explicitly estimating individual raters' centrality and severity under the same measurement framework allows researchers and practitioners to compensate for both of these rater effects simultaneously.

Some limitations of the present research should also be noted. The rating scale comprised only four categories. It goes without saying that the shorter the rating scale, the smaller the likelihood that centrality effects may manifest themselves. Jin and Wang (2018) analyzed a six-category rating scale and demonstrated a fairly strong impact of centrality effects on examinee proficiency rankings. It remains to be seen what the FM-SC will reveal about raters' central tendencies if the rating scale used in the TestDaF context would be longer. A six-category rating scale is currently under development.

A related limitation refers to the relatively small size of the examinee sample in Study 1 (N = 307) and Study 2 (N = 206), respectively. As Jin and Wang did in their study (N = 1,198), using a much larger examinee sample would have raised the likelihood of obtaining higher numbers of low-performing and high-performing examinees, respectively. This, in turn, would have affected the magnitude of the centrality effects' potential impact on examinee proficiency estimates and the differences between FM-SC and FM-S examinee rank orderings.

The present analyses were based on the rating scale instantiation of the FM-SC and FM-S, respectively. Different from Jin and Wang's (2018) partial credit model, the assumption, therefore, was that ratings on all criteria followed the same rating scale category structure. This choice was deliberate, given the small size of the examinee samples and the intended comparisons between the Bayesian estimation findings and the usual non-Bayesian facets analyses, which are mostly based on a rating scale model. In future studies, the rating scale and partial credit versions of the Bayesian FM-SC approach will be implemented and compared.

Finally, the relationships between centrality parameter estimates and three indirect statistical indices of rater centrality were studied building on real data sets. In each of these data sets, a range of factors may have contributed to the observed correlations (e.g., sample differences in the distribution of examinee proficiency levels or differences in the rater groups' mean severity level). Therefore, firm conclusions about the comparability of centrality estimates and indices cannot be drawn at this point. Simulation studies are needed, systematically varying potentially relevant factors and examining their impacts on the parameter estimates, the statistical indices, and their interrelationship.

Conclusion

The list of rater effects threatening rater-mediated assessments' validity and fairness is long (Myford & Wolfe, 2003, 2004; Saal et al., 1980; Wolfe & Song, 2016). Longer still is the list of statistical indices, psychometric models, and measurement approaches to detect these effects and examine their influence on the assessment outcomes (Eckes, 2015, 2017; Engelhard & Wind, 2018; Gwet, 2014). Jin and Wang's (2018) extended facets model (FM-SC) provides a rigorous psychometric method firmly grounded in the many-

facet Rasch measurement framework. The present research highlighted the FM-SC's potential to account for both rater severity and centrality, thus increasing the validity and fairness of the scores that raters assign to examinees. Through Bayesian MCMC estimation methods, the FM-SC implementation demonstrated the usability and utility of more complex and powerful facets models in applied assessment contexts. In these contexts, the FM-SC approach provides a significant advance for detecting and systematically analyzing centrality effects. This approach also enables the adjustment of ratings for between-rater centrality (and not only for between-rater severity) differences, much as Myford and Wolfe (2004) had called for so emphatically.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573. https://doi.org/10.1007/BF02293814

Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement Transactions*, *12*(3), 648-649. https://www.rasch.org/rmt/rmt1239.htm

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press. https://www.coe.int/en/web/portfolio/the-common-european-framework-of-reference-for-languages-learning-teaching-assessment-cefr-

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9). https://www.jstatsoft.org/article/view/v071i09

Denwood, M. J. (2019). *Package 'runjags'* (Version 2.0.4-6) [Computer software]. https://cran.r-project.org/web/packages/runjags/index.html

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197-221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2015). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments (2nd ed.). Peter Lang. https://doi.org10.3726/978-3-653-04844-5

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings - Part I [Editorial]. *Psychological Test and Assessment Modeling*, *59*(4), 443-452. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2017_20171218/03_Eckes.pdf

Eckes, T. (2018). Rater effects: Advances in item response modeling of human ratings - Part II [Editorial]. *Psychological Test and Assessment Modeling*, *60*(1), 29-32. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/2_Editorial__2018-03-10__1838.pdf

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Vol. 1. Fundamental techniques* (pp. 153-175). Routledge. https://doi.org/10.4324/9781315187815

Eckes, T. (2020). Rater-mediated listening assessment: A facets modeling approach to the analysis of raters' severity and accuracy when scoring responses to short-answer questions. *Psychological Test and Assessment Modeling*, *65*(4), 449-471. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-4/PTAM_4-2020_21900493_ebook_eckes_3.pdf

Eckes, T., & Althaus, H.-J. (2020). Language proficiency assessments in higher education admissions. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admission practices: An international perspective* (pp. 256-275). Cambridge University Press. https://doi.org/10.1017/9781108559607

Eckes, T., Müller-Karabil, A., & Zimmermann, S. (2016). Assessing writing. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 147-164). De Gruyter. https://doi.org/10.1515/9781614513827-012

Engelhard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191. https://doi.org/10.1207/s15324818ame0503_1

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261-287). Erlbaum.

Engelhard, G., & Wind, S. A. (2018). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. Routledge. https://doi.org/10.4324/9781315766829

Engelhard, G., & Wind, S. A. (2019). Introduction to the special issue on rater-mediated assessments [Editorial]. *Journal of Educational Measurement*, *56*(3), 475-477. https://doi.org/10.1111/jedm.12221

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*, 328-347. https://doi.org/10.1037/met0000059

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457-472. https://projecteuclid.org/download/pdf_1/euclid.ss/1177011136

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Guilford, J. P. (1936). *Psychometric methods*. McGraw-Hill.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Advanced Analytics.

Jackman, S. (2009). Bayesian analysis for the social sciences. Wiley.

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement, 74*, 116-138. https://doi.org/10.1177/0013164413498876

Jin, K.-Y., & Wang, W.-C. (2018). A new facets model for rater's centrality/extremity response style. *Journal of Educational Measurement*, *55*(4), 543-563. https://doi.org/10.1111/jedm.12191

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). Assessing performance: Designing, scoring, and validating performance tasks. Guilford.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kruschke, J. K. (2015). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd ed.). Academic Press/Elsevier.

Lane, S., & Iwatani, E. (2016). Design of performance assessments in education. In S. Lane, M. R. Raymond & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 274-293). Routledge.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Chapman & Hall/CRC.

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

Linacre, J. M. (2000). Comparing and choosing between "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). *Rasch Measurement Transactions*, *14*, 768. https://www.rasch.org/rmt/rmt143k.htm

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). JAM Press.

Linacre, J. M. (2006). Demarcating category intervals: Where are the category boundaries on the latent variable? *Rasch Measurement Transactions*, *19*, 1041-1043. https://www.rasch.org/rmt/rmt194f.htm

Linacre, J. M. (2020). Facets Rasch measurement computer program (Version 3.83) [Computer software]. Winsteps.com.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian analysis*. Chapman & Hall/CRC.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174. https://doi.org/10.1007/BF02296272

McNamara, T. F. (2000). *Language testing*. Oxford University Press.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: The role of measurement*. Oxford University Press.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176. https://doi.org/10.1177/014662169201600206

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

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

Norris, J., & Drackert, A. (2018). Test review: TestDaF. *Language Testing*, *35*(1), 149-157. https://doi.org/10.1177/0265532217715848

Plummer, M. (2017). *JAGS version 4.3.0 user manual*. https://web.sgh.waw.pl/~atoroj/ekonometria_bayesowska/jags_user_manual.pdf

R Core Team (2020). R: A language and environment for computing (Version 3.6.3) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods and implementation in R. *Psychological Test and Assessment Modeling*, *60*(1), 101-138. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/6_PTAM_IRMHR_Main__2018-03-13_1416.pdf

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151-1172. https://projecteuclid.org/download/pdf_1/euclid.aos/1176346785

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413-428. https://doi.org/10.1037/0033-2909.88.2.413

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*(4), 375-394. https://doi.org/10.1111/j.1745-3984.2005.00021.x

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*, 516-517. https://www.rasch.org/rmt/rmt103a.htm

Song, T., & Wolfe, E. W. (2015). *Distinguishing several rater effects with the Rasch model* [Paper presentation]. National Council of Measurement in Education Annual Meeting, Chicago, IL.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, *Series B*, *64*, 583-616. https://doi.org /10.1111/1467-9868.00353

Stafford, R. E., Wolfe, E. W., Casabianca, J. M., & Song, T. (2018). Detecting rater effects under rating designs with varying levels of missingness. *Journal of Applied Measurement*, *19*(3), 243-257. http://jampress.org/pubs.htm

Uto, M., & Ueno, M. (2020). A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika*, *47*(2), 469-496. https://doi.org/10.1007/s41237-020-00115-7

Wang, W.-C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296-318. https://doi.org/10.1177/0146621605276281

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

Weigle, S. C. (2012). Assessing writing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 218-224). Cambridge University Press.

Wind, S. A. (2019). Examining the impacts of rater effects in performance assessments. *Applied Psychological Measurement*, *43*(2), 159-171. https://doi.org/10.1177/0146621618789391

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161-192. https://doi.org/10.1177/0265532216686999

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35-51.

Wolfe, E. W., Chiu, C. W. T., & Myford, C. M. (2000). Detecting rater effects in simulated data with a multifaceted Rasch rating scale model. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147-164). Ablex.

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

Wolfe, E. W., Myford, C. M., Engelhard, G., & Manalo, J. R. (2007*). Monitoring reader performance and DRIFT in the AP English Literature and Composition examination using benchmark essays* (College Board Research Report No. 2007-2). College Board.

Wolfe, E. W., & Song, T. (2014). Rater effect comparability in local independence and rater bundle models. *Journal of Applied Measurement, 15*, 152-159. http://jampress.org/pubs.htm

Wolfe, E. W., & Song, T. (2015). Comparison of models and indices for detecting rater centrality. *Journal of Applied Measurement, 16*(3), 228-241. http://jampress.org/pubs.htm

Wolfe, E. W., & Song, T. (2016). Methods for monitoring and document rating quality. In H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common core standards, smarter-balanced, PARCC, and the nationwide testing movement* (pp. 107-142). Information Age.

Wolfe, E. W., & Wendler, C. (2020). Why should we care about human raters? [Editorial]. *Applied Measurement in Education*, *33*(3), 189-190. https://doi.org/10.1080/08957347.2020.1750407

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling*, *59*(4), 453-470. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2017_20171218/04_Wu.pdf

Appendix

Scale category frequencies for 18 raters in the Study 1 writing assessment

			b. TDN 3		TDN 3		TDN 4		TDN 5	
Rater	*N*	ω Est. (*SE*)	*n*	%	*n*	%	*n*	%	*n*	%
1	60	1.44 (.30)	0	0	1	1.7	27	45.0	32	53.3
17	135	1.28 (.12)	3	2.2	32	23.7	65	48.1	35	25.9
4	72	1.27 (.15)	3	4.2	24	33.3	35	48.6	10	13.9
9	141	1.24 (.11)	23	16.3	49	34.8	60	42.6	9	6.4
7	204	1.22 (.10)	5	2.5	39	19.1	98	48.0	62	30.4
3	123	1.11 (.12)	7	5.7	19	15.4	61	49.6	36	29.3
14	129	1.10 (.11)	16	12.4	52	40.3	48	37.2	13	10.1
6	102	1.10 (.10)	14	13.7	31	30.4	40	39.2	17	16.7
11	57	1.03 (.14)	7	12.3	23	40.4	16	28.1	11	19.3
13	123	1.00 (.11)	41	33.3	39	31.7	31	25.2	12	9.8
8	141	0.97 (.09)	28	19.9	39	27.7	51	36.2	23	16.3
12	132	0.96 (.10)	18	13.6	40	30.3	50	37.9	24	18.2
10	123	0.94 (.10)	22	17.9	40	32.5	41	33.3	20	16.3
2	72	0.87 (.13)	4	5.6	11	15.3	31	43.1	26	36.1
18	63	0.82 (.13)	7	11.1	18	28.6	18	28.6	20	31.7
15	84	0.76 (.11)	12	14.3	30	35.7	23	27.4	19	22.6
5	123	0.73 (.09)	31	25.2	38	30.9	31	25.2	23	18.7
16	60	0.69 (.14)	25	41.7	18	30.0	7	11.7	10	16.7

Note. The categories of the four-category TDN rating scale were *below TDN 3*, *TDN 3*, *TDN 4*, and *TDN 5* (higher categories indicate higher writing proficiency). Rater centrality estimates are shown for ease of reference. Raters are ordered by centrality estimates, from high (centrality) to low (extremity).

Scale category frequencies for 12 raters in the Study 2 writing assessment

			b. TDN 3		TDN 3		TDN 4		TDN 5	
Rater	*N*	ω Est. (*SE*)	*n*	%	*n*	%	*n*	%	*n*	%
1	270	1.32 (.08)	32	11.9	116	43.0	106	39.3	16	5.9
12	270	1.20 (.08)	45	16.7	116	43.0	81	30.0	28	10.4
10	270	1.15 (.07)	34	12.6	105	38.9	97	35.9	34	12.6
5	260	1.13 (.08)	28	10.8	101	38.8	102	39.2	29	11.2
3	270	1.04 (.06)	49	18.1	98	36.3	86	31.9	37	13.7
6	270	1.03 (.08)	21	7.8	69	25.6	115	42.6	65	24.1
7	270	1.02 (.07)	41	15.2	88	32.6	99	36.7	42	15.6
8	270	0.88 (.07)	56	20.7	95	35.2	84	31.1	35	13.0
2	270	0.88 (.06)	65	24.1	83	30.7	83	30.7	39	14.4
4	270	0.87 (.07)	51	18.9	99	36.7	83	30.7	37	13.7
9	270	0.86 (.07)	91	33.7	105	38.9	55	20.4	19	7.0
11	270	0.80 (.06)	53	19.6	91	33.7	76	28.1	50	18.5

Note. The categories of the four-category TDN rating scale were *below TDN 3*, *TDN 3*, *TDN 4*, and *TDN 5* (higher categories indicate higher writing proficiency). Rater centrality estimates are shown for ease of reference. Raters are ordered by centrality estimates, from high (centrality) to low (extremity).