MEASUREMENT: INTERDISCIPLINARY RESEARCH AND PERSPECTIVES
2022, VOL. 20, NO. 4, 228–247
https://doi.org/10.1080/15366367.2021.1972654

Detecting Rater Centrality Effects in Performance Assessments: A Model-Based Comparison of Centrality Indices

Kuan-Yu Jin[a] and Thomas Eckes[b]

[a]Assessment Technology and Research Division, Hong Kong Examinations and Assessment Authority; [b]TestDaF Institute, University of Bochum

ABSTRACT

Recent research on rater effects in performance assessments has increasingly focused on rater centrality, the tendency to assign scores clustering around the rating scale's middle categories. In the present paper, we adopted Jin and Wang's (2018) extended facets modeling approach and constructed a centrality continuum, ranging from raters exhibiting strong central tendencies to raters exhibiting strong tendencies in the opposite direction (extremity). In two simulation studies, we examined three model-based centrality detection indices (rater infit statistics, residual–expected correlations, and rater threshold *SD*s) as well as the raw-score *SD* in terms of their efficiency of reconstructing the true rater centrality rank ordering. Findings confirmed the superiority of the residual–expected correlation, rater threshold *SD*, and raw-score *SD* statistics, particularly when the examinee sample size was large and the number of scoring criteria was high. By contrast, the infit statistic results were much less consistent and, under conditions of large differences between criterion difficulties, suggested erroneous conclusions about raters' central tendencies. Analyzing real rating data from a large-scale speaking performance assessment confirmed that infit statistics are unsuitable for identifying raters' central tendencies. The discussion focuses on detecting centrality effects under different facets models and the indices' implications for rater monitoring and fair performance assessment.

KEYWORDS

Rater effects; rater centrality; facets models; performance assessment; many-facet Rasch measurement

Assessments in the social, behavioral, and health sciences often rely on human raters to evaluate examinees' performance on a given task or item (Kane et al., 1999; Lane & Stone, 2006). The tasks may range from limited production tasks like short-answer questions to extended production tasks, prompting examinees to write an essay, deliver a speech, or provide work samples (Carr, 2011; Johnson et al., 2009). Whatever the exact nature of the tasks, human ratings of examinee performance have a role in making decisions, including high-stakes decisions on university admission, graduation, certification, immigration, or staff recruitment. Moreover, in developing and implementing automated scoring systems, human ratings are considered the "gold standard" for determining the accuracy of the scores these systems produce (Powers et al., 2015; Williamson et al., 2012; Wolfe, 2020). Therefore, it is essential to ensure that the assessments conform to high psychometric quality standards, particularly regarding their reliability, validity, and fairness (Engelhard & Wind, 2018; Lane & DePascale, 2016; Penfield, 2016).

However, the level of rating quality achievable in an assessment is generally limited by raters' judgmental and decision-making processes, in particular, their perception of performance features, their interpretation and use of the scoring rubric, and their more or less implicit response biases (Bejar, 2012; Eckes, 2012; Glazer & Wolfe, 2020; Lane, 2019; Myers et al., 2020). Together referred to as rater

CONTACT Thomas Eckes thomas.eckes@gast.de TestDaF Institute, University of Bochum, Universitätsstr. 134, 44799 Bochum, Germany

© 2021 Taylor & Francis Group, LLC

effects, these varied and mostly subtle rater influences have long been known to threaten the validity of the assessment outcomes. There has been a similarly long tradition of research on psychometric models and statistical indices for detecting these effects (Myford & Wolfe, 2003, 2004; Saal et al., 1980; Wolfe & Song, 2016).

Much of this research has been concerned primarily with rater severity (Engelhard, 1992; Lunz et al., 1990, 1996; McNamara, 1996). Severity effects manifest themselves through decreases in the *average level* of the scores assigned to examinees, decreases not expected given the quality of their performances. With its opposite tendency (i.e., leniency), rater severity is arguably the most pervasive and detrimental effect. More recently, the researchers' attention has increasingly been directed toward another rater effect – centrality (Falk & Cai, 2016; Jin & Wang, 2014, 2018; Uto & Ueno, 2020; Wolfe & Song, 2015, 2016). Rater centrality effects manifest themselves through decreases in the *dispersion* of the scores assigned to examinees, decreases not expected given the variability of their performances. That is, raters subject to centrality tend to assign scores unduly clustering around the rating scale's middle category (or categories). With its opposite tendency (i.e., extremity or extreme response style), centrality (much like severity) covers varying degrees of negative influences on an assessment's validity and fairness (Myford & Wolfe, 2003, 2004; Wolfe & Song, 2016).

The present research focuses on centrality effects, specifically, how to detect, measure, and control these effects in the context of large-scale performance assessments. In simulation and real-data studies, we examined the usefulness of several statistical indices proposed in the literature for detecting centrality effects (Myford & Wolfe, 2004; Wolfe & Song, 2015, 2016; Wu, 2017).

For this purpose, we built on the extended facets modeling approach proposed by Jin and Wang (2018). This approach allowed us to construct a continuum of centrality, ranging from raters exhibiting strong central tendencies to raters exhibiting strong tendencies in the opposite direction (extremity). Also, we constructed a separate continuum of severity for the same group of raters, distinguishing between severe and lenient raters. We then compared the raters' location on the centrality continuum to their centrality values provided by the different statistical indices, systematically varying several characteristics typical of performance assessments. Finally, we probed these indices' practical utility using real rating data from a large-scale speaking assessment.

Indirect modeling approaches to rater centrality

Studies adopting a latent trait or item response theory (IRT) modeling approach to rater centrality have typically followed an indirect, two-step procedure: In the first step, the rating data are analyzed to estimate a single rater parameter, that is, rater severity. In the second step, a statistical index is computed based on the findings to capture individual raters' central (or extreme response) tendency (Wolfe & Song, 2015, 2016). By contrast, in this paper, we advocate a direct approach, distinguished by simultaneously estimating two rater parameters: rater severity and centrality (we discuss the direct approach in more detail later). As for the indirect approach, Table 1 summarizes the key features of three model-based centrality indices that have been in use since the early 1990s.

The first rater centrality index, the *rater infit statistic* (MS_W), belongs to the family of residual fit statistics (DeMars, 2017; Wright & Masters, 1982; Wu & Adams, 2013). In the context of human ratings, these statistics indicate the extent to which scores provided by a given rater match the scores expected under a particular psychometric model.

In a three-facet assessment situation where J raters assign scores to N examinees on I tasks (or items, criteria) using a rating scale with $m + 1$ categories, that is, $k = 0, \ldots, m$, the following many-facet Rasch measurement or *facets model*, for short, an extension of the rating scale model (Andrich, 1978), may be used (Eckes, 2015; Engelhard & Wind, 2018; Linacre, 1989):

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

Table 1. Overview of model-based rater centrality indices.

Centrality index	Values[a]	Facets model[b]	References[c]
Weighted mean-square residual fit (rater infit, MS_W)	Low (overfit, e.g., $MS_W < 0.75$)	Rating scale model (RSM; Equation (1))	Engelhard (1992, 1994); Myford and Wolfe (2004); Smith (1996); Wolfe et al. (2000, 2007)
Residual–expected correlation ($r_{\text{res,exp}}$)	Negative (e.g., $r_{\text{res,exp}} < -.30$)	Rating scale model (RSM; Equation (1))	Wolfe (2004); Wolfe and McVay (2012); Wolfe and Song (2014, 2015)
Standard deviation of rater threshold estimates ($SD(\tau_{jk})$)	High	Rater-related partial credit model (PCM; Equation (2))	Myford and Wolfe (2004); Song and Wolfe (2015); Stafford et al. (2018); Wu (2017)

[a]Entries indicate when raters would be identified as manifesting centrality under each index.
[b]Model version typically used for computing each centrality index.
[c]Sample studies that examined one or more centrality indices.

where p_{nijk} is the probability of examinee n receiving a rating of k from rater j on task i, $p_{nij(k-1)}$ is the probability of examinee n receiving a rating of k – 1 from rater j on task i, θ_n is the ability of examinee n, β_i is the difficulty of task 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.

Based on the three-facet rating scale model (Equation (1)), the rater infit statistic is computed as the average of the squared standardized residuals over all examinees and tasks involved in producing that rater's scores, each squared standardized residual weighted by its variance or statistical information (Eckes, 2015; Engelhard & Wind, 2018). Infit statistics have an expected value of 1.0 and range from 0 to plus infinity (Linacre, 2002; Myford & Wolfe, 2003). Thus, fit values greater than 1.0 indicate more variation than expected in the ratings (reduced predictability). By contrast, fit values less than 1.0 indicate less variation than expected; that is, the ratings tend to be muted or overpredictable; 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 1.0, in particular, less than 0.70 or 0.75 (e.g., Engelhard, 1994). For this reason, rater overfit has been suggested as a potential indicator of rater centrality. More recently, Wind and Jones (2019) found that average infit statistics for raters simulated to exhibit centrality (the "effect" or "central" raters) were lower than the average infit statistics for the "no-effect" raters; the infit values for the central raters ranged from 0.35 to 0.48, irrespective of the rating design used.

However, evidence has also accumulated that fit statistics may be sensitive to various other rater effects, such as halo error, inaccuracy or randomness, and range of restriction (Myford & Wolfe, 2004; Wolfe, 2004; Wolfe et al., 2000, 2007), calling into question any straightforward interpretation of rater fit statistics in terms of centrality effects. For example, Wolfe et al. (2000) observed rater fit statistics greater than 1.0 for raters simulated to exhibit centrality – a finding in stark contrast to Wind and Jones (2019).

Myford and Wolfe (2004) suggested a possible explanation for observing rater fit values greater than 1.0 in the presence of centrality effects. They created a hypothetical example where a rater scored a single examinee on a set of 10 traits differing widely in difficulty. When this rater assigns average scores on each trait, thus exhibiting centrality, the observed scores will show no variation and mismatch the expected scores. As a result, the infit statistic will be greater than the expected value of 1.0 in this case. Conversely, when another rater provides accurate scores on each trait, faithfully representing each trait's difficulty, the observed scores will vary widely and perfectly match the expected scores, resulting in an infit value close to zero. Therefore, in this example, the usual way of interpreting the infit statistic is completely disrupted.

Though Myford and Wolfe's (2004) example may not be typical of the conditions found in operational performance assessments, as a safeguard against incorrect inferences about a given rater's centrality, they recommended that "the researcher carefully examine vectors of observed ratings for all overfitting or misfitting raters before concluding that they are exhibiting a central tendency effect" (Myford & Wolfe, 2004, p. 203). Thus, in one of our simulation studies, we will examine when to expect rater infit statistics greater than 1.0 even if raters demonstrate centrality.

Wolfe (2004) proposed another centrality index, the *residual–expected correlation statistic* ($r_{\text{res,exp}}$). The rationale underlying this index is as follows: When rater j exhibits a centrality effect, the scores this rater assigns to high-proficient examinees are lower than the expected ratings; hence, the residuals will be large and negative. Conversely, the scores assigned to low-proficient examinees are higher than the expected ratings; the residuals will be large and positive in this case. As a result, the Pearson correlation between residual and expected scores 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 build on the facets model's rating scale version (Equation (1)). A third centrality index builds on a particular version of the partial credit model (Masters, 1982). In the present facets context, this 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 contrast to the previous model (Equation (1)), raters are no longer assumed to share the same rating scale structure. Rather, the rating scale for each rater is modeled to have its own category structure.

Considering a partial credit model like this, Myford and Wolfe (2004; see also Wolfe & Song, 2015; Wu, 2017) suggested using the *standard deviation of the rater threshold parameter estimates* ($SD(\tau_{jk})$), or *rater threshold SD*, for short, to detect central tendencies. Their proposal rests on the following reasoning: When raters exhibit centrality effects, they tend to include a wide range of examinee proficiency levels in the rating scale's middle category (or categories). In this case, the lower thresholds will drop and the higher thresholds will rise. Therefore, 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.

Model-based centrality indices share the feature of utilizing the complete set of ratings for parameter estimation. Alternatively, there has been some tradition of using observed score frequencies to indicate central tendencies (Johnson et al., 2009). Specifically, one may compute, separately for each rater, the standard deviation of the scores assigned, that is, the observed or raw score *SD* (SD_{obs}), with smaller SD_{obs} indicating greater centrality. However, the use and interpretation of this index are limited because it lacks a suitable basis for evaluation in most practical applications (Wolfe, 2020; Wolfe & Song, 2016). For example, when comparing individual raters' raw scores to consensus scores assigned by a group of expert raters, the set of jointly scored performances is typically small and possibly biased along the rating scale in terms of preselected performance levels. Nonetheless, in our simulation and real-data studies, we will consider the SD_{obs} index because it is usually the only centrality index available to practitioners without further psychometric analysis.

Direct rater centrality modeling

Each facets model shown in Equations (1) and (2) includes a parameter (α_j) directly modeling the severity of rater j. Consequently, estimates of examinee proficiency are corrected for between-rater severity differences in much the same way as they are corrected for between-item or between-task difficulty differences (for example, when examinees are free to choose which item or task to work on). However, for rater centrality, the situation is fundamentally different. Facets models currently in use do not include a rater centrality parameter and, therefore, do not provide rater centrality estimates, nor do they provide examinee proficiency estimates that compensate for between-rater centrality differences. Adopting Jin and Wang (2018) terminology, these models are called *facets models–severity only* or *FM-S*, for short.

Extending these models, Jin and Wang (2018) proposed a facets model incorporating a rater centrality parameter. Building on the same three-facet situation as before, this model is defined as follows:

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

where all parameters are as in Equations (1) and (2) except for the $\omega_j\tau_{ik}$ term. In this term, the parameter ω_j (with $\omega_j > 0$) is a weight parameter representing the centrality of rater *j*. Also, the parameter τ_{ik} denotes the threshold for a particular task; that is, the difficulty of receiving a rating of *k* relative to *k* – 1 when considering task *i*. Unlike the previous models (Equations (1) and (2)), it is assumed that the rating scale structure varies from task to task. Hence, Equation (3) specifies a three-facet task-related partial credit model (PCM) accounting for rater severity and centrality. Alternatively, a rating scale (RSM) version of this model may be specified by replacing, in Equation (3), the task-related threshold parameter τ_{ik} by τ_k. Following Jin and Wang (2018), facets model extensions of this kind are subsumed under the term *facets model–severity and centrality* or *FM-SC*, for short.

Given the parameterization shown in Equation (3), the higher the value of ω_j, the more rater *j* tends to overuse the middle categories of the rating scale, exhibiting a centrality effect; conversely, the lower the value of ω_j, the more that rater tends to overuse the extreme categories of the rating scale, exhibiting an extremity effect; when $\omega_j = 1$ for $j = 1, \ldots, J$, the FM-SC reduces to the FM-S.

Jin and Wang (2018) estimated the model parameters specified in the FM-SC (Equation (3)), building on a Bayesian approach (Levy & Mislevy, 2016). Following this approach, Eckes and Jin (2021a, 2021b) applied FM-SC's rating scale and partial credit versions to different real datasets taken from writing and speaking performance assessments administered in higher education admissions contexts. In the first study (Eckes & Jin, 2021b), two datasets from writing assessments utilizing an analytic scoring rubric provided the input to the FM-SC rating scale version; the datasets contained scores on three or nine criteria, respectively. In the second study (Eckes & Jin, 2021a), two different datasets from writing and speaking assessments utilizing a holistic scoring rubric provided the input to the rating scale and partial credit FM-SC versions; the datasets contained scores on two writing tasks and six speaking tasks, respectively.

Across studies and datasets, raters showed significant differences in their severity and centrality. What is more, the centrality effects had a practically relevant impact on examinee rank orderings. For example, in the Eckes and Jin (2021a) study, even though examinee proficiency measures estimated under the FM-SC and the FM-S were highly correlated, the resulting rank orders of examinees, sorted from high to low proficiency, differed on average by about five (writing) or three ranks (speaking).

In both studies, the model-based indices of rater centrality discussed above were derived from FM-S analyses run in parallel with the FM-SC. The authors correlated rater infit statistics, residual–expected correlations, and rater threshold *SD*s with each other and with the ω estimates. The correlational patterns found were mostly consistent across the four datasets. With one exception, the correlations between ω estimates and the indices were moderately high to strong, statistically highly significant, and in the expected direction. Thus, higher central tendencies estimated under the FM-SC were associated with (a) lower infit values (overfit), (b) lower (typically negative) correlations between residual and expected scores, and (c) higher standard deviations of the Rasch-Andrich threshold estimates. The exception concerned the correlation between rater infit statistics and ω estimates, *r*(12) = –.41, *ns*, in the second dataset, first study (Eckes & Jin, 2021a).

Regarding the indices' intercorrelations, those involving rater infit showed considerable inconsistency across studies and datasets. For example, the correlations for the residual–expected correlation statistic ranged from $r(30) = .71$, $p < .01$, second dataset, second study (Eckes & Jin, 2021a), to $r(12) = -.25$, *ns*, second dataset, first study (Eckes & Jin, 2021b). These inconsistent correlations are reminiscent of Myford and Wolfe's (2004) recommendation to take a closer look at the misfitting and overfitting raters' scoring patterns before drawing conclusions about their central tendencies.

In the Eckes and Jin (2021a, 2021b) studies, the relationships between centrality parameter estimates and the statistical indices were examined using real datasets. Therefore, a range of factors may have contributed to the correlations' observed magnitude and direction. For example, the examinee sample size and the number of tasks or scoring criteria may have affected the statistical indices in various, unknown ways. Consequently, firm conclusions about each of the indices' suitability for detecting centrality effects could not be drawn on that basis. In situations like this, simulation studies, where some factors are systematically varied, and others are deliberately held constant, may yield more valid insights.

Several studies have adopted the simulation approach, examining rater effects using various statistical indices and determining the rater effects' impact on the measurement quality of assessment outcomes (e.g., Song & Wolfe, 2015; Stafford et al., 2018; Wind, 2019, 2020; Wind & Jones, 2019; Wolfe & Song, 2015). Typically, these studies have used psychometric models belonging to the FM-S class (see above) to generate data under different rater-effect conditions; that is, the rating data were simulated to exhibit certain kinds of rater effects.

For example, Wolfe and Song (2015) used RSM and PCM versions of the FM-S class as well as generalized models (Muraki, 1992), incorporating a slope (discrimination) parameter interpreted as a rater centrality index. Besides distinguishing between "normal" and "central" raters, the simulation design variables included the number of rating scale categories (three or five), 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). Results showed that the residual–expected correlation index performed better than rater infit statistics, rater threshold *SD*s, and rater slope, providing low Type I error rates (i.e., incorrectly flagging raters who did not exhibit a centrality effect) and low Type II error rates (i.e., not flagging raters who exhibited a centrality effect).

In another simulation study, Song and Wolfe (2015) used rating data containing multiple rater effects. They found that Rasch measures of rater location, rater threshold *SD*, and correlations between observed ratings and estimated examinee proficiencies performed well in detecting rater severity, centrality, and inaccuracy, respectively. Similarly, Wind and Guo (2019) studied the combined effects of differential rater functioning (DRF) and rater misfit on rater effect detection. The findings confirmed that rater fit statistics, rater discrimination estimates, and commonly-used DRF statistics were sufficiently sensitive to these rater effects.

In the present research, different from Wolfe and Song (2015) and related simulation studies, we built on the direct FM-SC modeling approach (Jin & Wang, 2018). We sampled true rater centrality values under conditions typically found in operational large-scale performance assessments. In such assessments, ratings are often missing by design; that is, the assessments use incomplete rating designs where not every rater scores every examinee's performance (Eckes, 2015; Wind & Jones, 2019). The sampled centrality values were used to study the statistical indices' suitability for detecting rater centrality effects.

Research questions

Against the background of the FM-SC modeling approach to the study of rater centrality effects, the present research aimed to answer the following three questions:

(1) How do statistical indices, particularly rater infit statistics, residual–expected correlations, rater threshold *SD*s, and observed *SD*s, relate to rater centrality values sampled from a centrality distribution according to the FM-SC (Equation (3))? The simulation design included two factors: (a) examinee sample size and (b) the number of scoring criteria. Hence, the question becomes: What is the sensitivity of each of these indices to rater centrality differences under varying numbers of examinees and criteria?

(2) Under what conditions will rater infit values be greater than 1.0 when raters exhibit centrality? How do rater infit statistics, residual–expected correlations, rater threshold *SD*s, and observed *SD*s, behave when scoring criterion difficulties differ in range? This research question explicitly addressed Myford and Wolfe's (2004) suggestion that the trait (or, in our case, criterion) difficulty range influences the magnitude of rater infit statistics and severely limits their usefulness for detecting centrality effects.

(3) How do statistical centrality indices relate to each other in a large-scale assessment context, where raters used rating scales differing in length to evaluate examinee performance on a set of performance tasks?

We addressed the first two research questions using specifically designed simulation studies. The third question was examined reanalyzing real data from a speaking performance assessment.

Simulation study 1: number of examinees and criteria

Design

This simulation study aimed to answer the first research question: What is the impact of the number of examinees and criteria, respectively, on the correlation between the statistical indices and ω as introduced in Equation (3)? In the present study, ω represented the true magnitude of rater centrality; therefore, this is tantamount to the question: What is the indices' sensitivity for detecting centrality effects under systematically varied conditions?

Building on previous rater effect studies (e.g., Jin & Wang, 2018; Wind & Sebok-Syer, 2019), we created a simulation design involving two factors: (a) the number of examinees, with two levels (500 or 1,000 examinees), and (b) the number of scoring criteria, with three levels (two, three, or four criteria). The criterion difficulties were set as follows: –0.25 and 0.25 (two criteria), –0.5, 0, and 0.5 (three criteria), and –0.75, –0.25, 0.25, and 0.75 (four criteria).

We held the following characteristics constant across the six combinations of factor levels: Ratings were assigned by 25 raters using a five-category scale (from 0 = *low* to 4 = *high*) with thresholds fixed at –0.75, –0.25, 0.25, and 0.75. Each examinee provided a single performance, and two raters scored

Table 2. Spiral rating design used in simulation study 1.

Examinee group	Rater 1	Rater 2	Rater 3	. . .	Rater 24	Rater 25
Group 1	✓	✓				
Group 2		✓	✓			
. . .				. . .		
Group 24					✓	✓
Group 25	✓					✓

Each ✓ designates observed scores (ratings). Empty cells indicate missing ratings.

each performance, following an incomplete, spiral rating design (Jin & Wang, 2018; see also Eckes, 2015; Wind & Jones, 2019). Examinees were divided into 25 equally sized groups. As illustrated in Table 2, Group 1 was scored by Raters 1 and 2; Group 2 was scored by Raters 2 and 3; and so on.

Each rater in the 500-examinees condition assigned ratings to 40 performances. Hence, for a given rater in this condition, the number of ratings varied between 80 (two criteria) and 160 (four criteria). In the 1,000-examinees condition, each rater assigned ratings to 80 performances; thus, the number of ratings per rater was much higher, varying between 160 (two criteria) and 320 (four criteria).

Based on the FM-SC rating scale version, rater severity values were sampled from a normal distribution with mean 0 and variance 0.25; also, rater centrality values (i.e., true ω values) were sampled from a lognormal distribution with mean 0 and variance 0.25 (Levy & Mislevy, 2016; Lunn et al., 2013). Table 3 presents the resulting values for raters' severity and centrality, ordered by centrality from high to low. The severity values ranged from 1.09 (Rater 23) to −1.21 (Rater 7), with $M = 0$ ($SD = 0.51$); the centrality values ranged from 3.69 (Rater 6) to 0.17 (Rater 22), with $M = 1.31$

Table 3. Rater severity and centrality values used in the simulation studies.

Rater	Severity (α_j)	Centrality (ω_j)
6	−0.60	3.69
9	0.69	3.21
14	0.15	2.45
11	−0.09	2.30
19	0.16	2.21
4	−0.13	2.18
2	0.06	2.06
25	−0.47	1.71
18	0.00	1.65
16	0.63	1.16
10	−0.33	1.04
3	0.46	1.00
23	1.09	0.97
13	0.20	0.97
15	0.07	0.95
17	−0.34	0.90
21	−0.05	0.89
7	−1.21	0.76
24	0.38	0.58
1	−0.57	0.53
12	−0.20	0.52
8	0.78	0.33
20	0.10	0.33
5	−0.72	0.27
22	−0.05	0.17

Raters are ordered by centrality values, from high (centrality) to low (extremity).

Table 4. Design for simulation study 1.

Factors	Level(s)	Parameter setting
Manipulated		
Number of examinees	(1) 500	$\theta_n \sim N(0, 1)$
	(2) 1,000	$\theta_n \sim N(0, 1)$
Number of criteria	(1) 2	$\beta_1 = -0.25$, $\beta_2 = 0.25$
	(2) 3	$\beta_1 = -0.5$, $\beta_2 = 0$, $\beta_3 = 0.5$
	(3) 4	$\beta_1 = -0.75$, $\beta_2 = -0.25$, $\beta_3 = 0.25$, $\beta_4 = 0.75$
Held constant		
Number of scale categories	5	$\tau_1 = -0.75$, $\tau_2 = -0.25$, $\tau_3 = 0.25$, $\tau_4 = 0.75$
Number of raters	25	$\alpha_j \sim N(0, 0.25)$, $\omega_j \sim lognormal(0, 0.25)$

(*SD* = 0.94). In both simulation studies, severity and centrality values were held constant across conditions. Table 4 summarizes our Simulation Study 1 design in terms of the factors, levels, and parameter settings used.

Under each unique combination of factors (simulation conditions), we generated 100 replications. We computed rater infit statistics (MS_W) and residual–expected correlations ($r_{\text{res,exp}}$) for each replication following the rating scale FM-S specified in Equation (1). Likewise, building on the rater-related partial credit facets model (Equation (2)), we computed rater threshold *SD* statistics ($SD(\tau_{jk})$). We ran the rating scale and partial credit FM-S analyses using the R package TAM (Test Analysis Modules; Robitzsch et al., 2020) with marginal maximum likelihood estimation (MMLE).[1] Finally, we also computed the raw score *SD* statistic (SD_{obs}) to compare with model-based centrality indices.

Results

Table 5 presents the findings from the correlation analyses. As expected, the correlations for MS_W, $r_{\text{res,exp}}$, and SD_{obs} were consistently negative and the correlations for $SD(\tau_{jk})$ were consistently positive. Overall, the mean correlations were strong, and they were stronger still for the large sample condition. The number of criteria involved seemed to have less impact on the correlations' magnitude, particularly for the infit index.

Figure 1 displays the corresponding boxplots for the 500-examinees condition (upper panel) and the 1,000-examinees condition (lower panel). For ease of presentation, along the vertical dimension, the magnitude of the correlations between ω and MS_W, $r_{\text{res,exp}}$, and SD_{obs}, respectively, is shown in absolute values.

The boxplots demonstrate that the rater threshold *SD* index comes with some outlying or extreme correlations when the number of criteria is low (i.e., two or three criteria). This finding contrasts with the other indices, especially with the $r_{\text{res,exp}}$ index. However, when the scoring rubric includes four criteria, the rater threshold *SD* index manifests the overall smallest dispersion of correlations (see also the correlation *SD* values in Table 5). Of course, the number of criteria used when scoring examinee performances restricts the possible range of threshold values. This (trivial) dependency should be considered when interpreting rater threshold *SD* index values as evidence of rater centrality for a given real data set.

Table 5. Means and standard deviations of correlations between rater centrality indices and centrality parameter ω_j in simulation study 1.

Statistic	500 examinees			1,000 examinees		
	2 criteria	3 criteria	4 criteria	2 criteria	3 criteria	4 criteria
			Infit (MS_W)			
M	−.81	−.84	−.81	−.88	−.90	−.89
SD	0.05	0.05	0.05	0.03	0.03	0.03
Mdn	−.82	−.84	−.82	−.88	−.90	−.89
			Residual–expected correlation ($r_{\text{res,exp}}$)			
M	−.87	−.91	−.93	−.91	−.93	−.95
SD	0.04	0.02	0.02	0.02	0.02	0.01
Mdn	−.87	−.91	−.93	−.91	−.94	−.95
			Standard deviation of rater thresholds ($SD(\tau_{jk})$)			
M	.83	.92	.96	.95	.97	.98
SD	0.13	0.09	0.01	0.06	0.01	0.01
Mdn	.90	.95	.96	.96	.97	.98
			Standard deviation of observed scores (SD_{obs})			
M	−.84	−.87	−.88	−.88	−.89	−.90
SD	0.05	0.04	0.04	0.03	0.03	0.02
Mdn	−.84	−.87	−.88	−.88	−.90	−.91

Under each condition of the 2 (examinee sample size) × 3 (number of scoring criteria) factorial design, correlations were computed based on 100 replications.

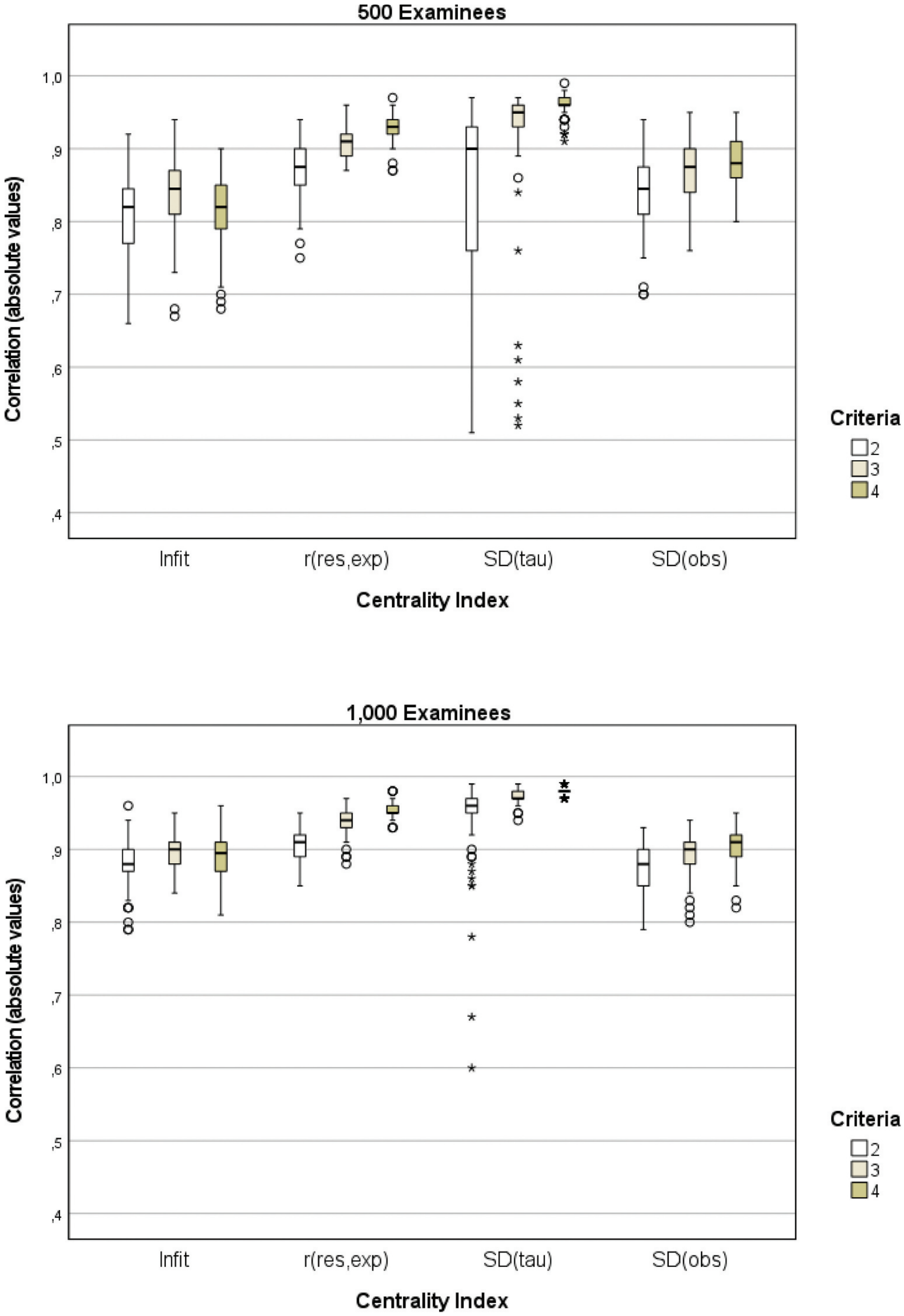

Figure 1. Correlations (absolute values) between four centrality indices and centrality parameter ω_j across different numbers of scoring criteria and two examinee samples (simulation study 1).

Simulation study 2: criterion difficulty differences

Design

This simulation study aimed to answer the second research question: What is the impact of the criterion difficulties' range on the correlation between the statistical indices and ω? In particular, we examined the assumption that the rater infit index fails to detect centrality effects when the criterion difficulties are widely dispersed. Therefore, this study also explained the puzzling observation of infit values greater than 1 when the data demonstrate centrality effects, as sometimes reported in the literature (Myford & Wolfe, 2004; Wolfe et al., 2000).

For answering this research question, we created a simulation design involving a single factor: The amount of dispersion between the criterion difficulties, with three range levels (low, medium, and high). We considered four criteria. Under the low-level condition, the difficulty of the first two criteria was fixed at −1, and the difficulty of the other two criteria was fixed at 1. The medium- and high-level conditions were defined by difficulties fixed at −1.5 vs. 1.5 and −2 vs. 2. All other settings were identical to Simulation Study 1, except for the examinee sample size; that is, we considered the 1,000-examinees condition only. Table 6 summarizes our Simulation Study 2 design in terms of the factors, levels, and parameter settings used.

Also, under each of the three range level conditions, we conducted a total of 100 replications. For each replication, we computed the model-based statistical indices using TAM (Robitzsch et al., 2020) and, finally, correlated the resulting index values and the raw score *SD* statistics with rater centrality (ω) values.

Table 6. Design for simulation study 2.

Factors	Level(s)	Parameter setting
Manipulated		
Range of criterion difficulties	(1) Low	$\beta_1 = -1, \beta_2 = -1, \beta_3 = 1, \beta_4 = 1$
	(2) Medium	$\beta_1 = -1.5, \beta_2 = -1.5, \beta_3 = 1.5, \beta_4 = 1.5$
	(3) High	$\beta_1 = -2, \beta_2 = -2, \beta_3 = 2, \beta_4 = 2$
Held constant		
Number of examinees	1,000	$\theta_n \sim N(0, 1)$
Number of score categories	5	$\tau_1 = -0.75, \tau_2 = -0.25, \tau_3 = 0.25, \tau_4 = 0.75$
Number of raters	25	$\alpha_j \sim N(0, 0.25), \omega_j \sim lognormal(0, 0.25)$

Table 7. Means and standard deviations of correlations between centrality indices and and centrality parameter ω_j in simulation study 2.

Statistic	Range of criterion difficulties		
	Low	Medium	High
		Infit (MS_W)	
M	−.71	−.04	.60
SD	0.08	0.17	0.12
Mdn	−.73	−.03	.61
		Residual–expected correlation ($r_{\text{res,exp}}$)	
M	−.98	−.98	−.98
SD	0.01	0.01	0.01
Mdn	−.98	−.98	−.98
		Standard deviation of rater thresholds ($SD(\tau_{jk})$)	
M	.98	.98	.98
SD	0.01	0.01	0.01
Mdn	.98	.98	.98
		Standard deviation of observed scores (SD_{obs})	
M	−.93	−.95	−.96
SD	0.02	0.01	0.01
Mdn	−.93	−.95	−.96

Under each of the three criterion difficulty ranges, correlations were computed based on 100 replications (examinee sample size was 1,000; number of scoring criteria was 4).

Results

The mean correlations (and medians) presented in Table 7 are again consistently negative for $r_{\text{res,exp}}$ and SD_{obs}, and consistently positive for $SD(\tau_{jk})$; also, the absolute values of correlation statistics are close to unity for $r_{\text{res,exp}}$ and $SD(\tau_{jk})$.

The MS_W index correlations demonstrate a completely different pattern: The mean correlations increase from −.71 under the low-level condition, a near-zero value under the medium-range level, and a positive value of .60 under the high-range level. Thus, all other things being equal, the more the criteria differ in their difficulties, the higher the infit values. This finding implies a complete reversal of this index's intended interpretation. Under such conditions, rater centrality would no longer be indicated by smaller MS_W values, particularly values much smaller than the expected value (i.e., 1.0); instead, centrality effects would be indicated by higher MS_W values, particularly values much greater than the expectation. Also, the relatively high correlation *SD* values show that correlations vary considerably within each of the three simulation conditions. By contrast, the corresponding *SD*s for the other indices stay close to zero.

The boxplots displayed in Figure 2 illustrate how much the correlations differ between centrality indices across range level conditions. The $r_{\text{res,exp}}$, $SD(\tau_{jk})$, and SD_{obs} indices provide highly consistent conclusions regarding the presence of rater centrality effects, irrespective of the extent to which the criterion difficulties differ from each other. The situation for the MS_W index is strikingly different: The correlations' strength and direction depend heavily on the range of criterion difficulties. Moreover, within each of the three range levels, the correlations show a great deal of variability.

Figure 3 displays the infit distributions for five selected raters under each criterion difficulty range for illustrative purposes, demonstrating that high centrality may be associated with infit values much greater than the expected value of 1.0. In the simulation, of all the 25 raters, Rater 6 had the strongest tendency toward centrality (ω = 3.69, see Table 3). Under the low-range condition, this rater's mean infit value stayed somewhat below 1.0 (M = 0.90, SD = 0.07, Mdn = 0.90). However, under the medium-range condition, the vast majority of values fell above 1.0 (M = 1.17, SD = 0.09, Mdn = 1.17).

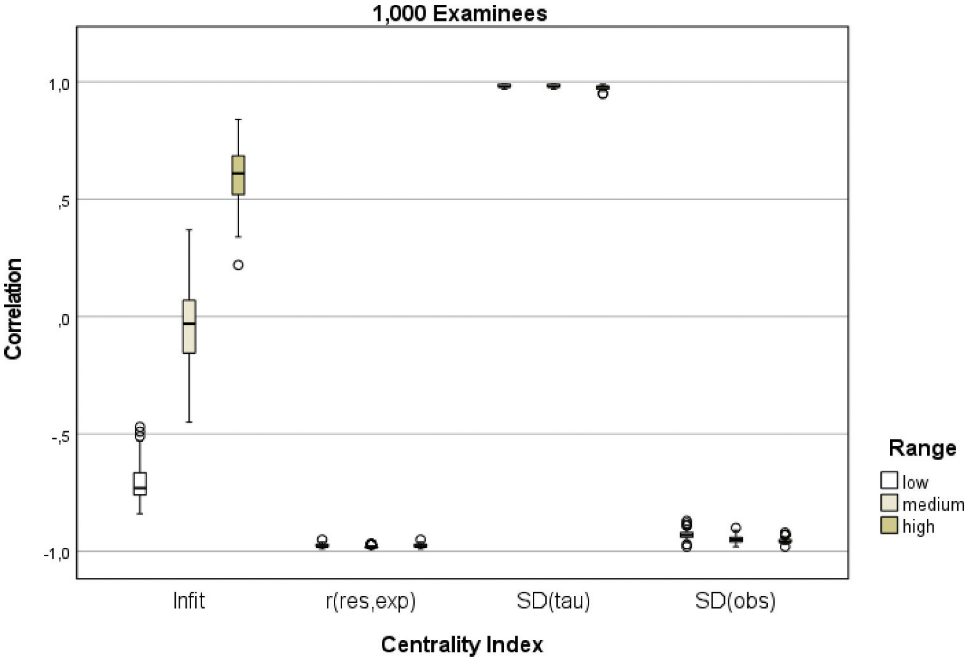

Figure 2. Correlations between four centrality indices and centrality parameter ω_j across three levels of criterion difficulty range (simulation study 2).

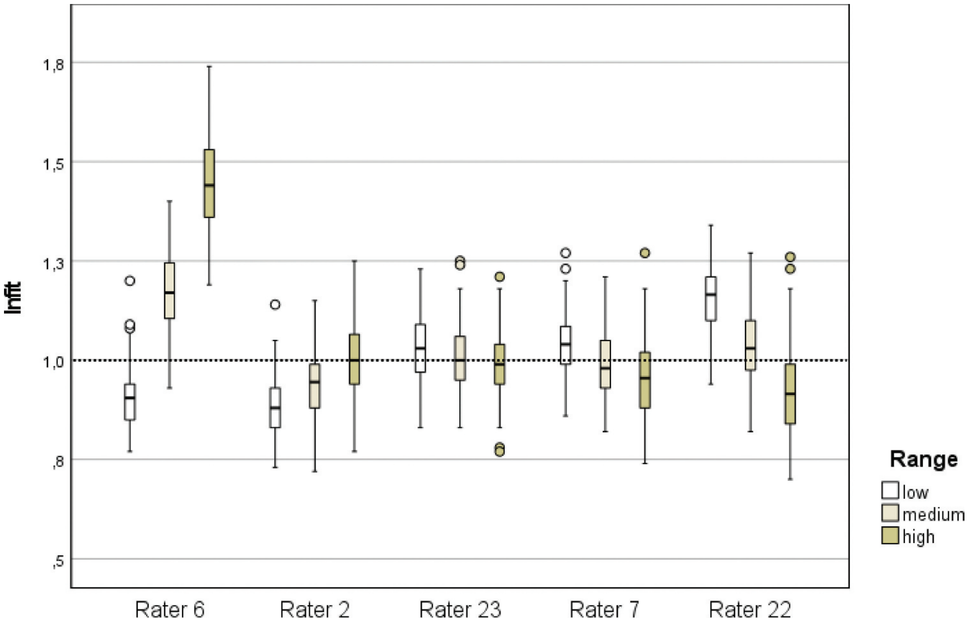

Figure 3. Infit values for five raters and three levels of criterion difficulty range (simulation study 2).

Finally, under the high-range condition, all infit values stayed well above 1.0, with some 30% of these values greater than 1.5 (M = 1.45, SD = 0.12, Mdn = 1.44). There is a similar progression of infit values across conditions for another rater with heightened centrality: Rater 2 (ω = 2.06).

Illustrating the converse case, Figure 3 also shows the infit values for two raters, Rater 7 (ω = 0.76) and Rater 22 (ω = 0.17), tending toward the scale's extreme categories. The mean infit values monotonically *decrease* across conditions for both raters, even falling below 1.0 under the high-range condition. This finding suggests that low infit values indicate extremity–quite in contrast to the commonly intended interpretation of the rater infit statistic.

Real-data study: TestDaF speaking assessment

Instrument and procedure

The Test of German as a Foreign Language (TestDaF, *Test Deutsch als Fremdsprache*) is officially recognized as a language exam for international students applying for entry to higher education institutions in Germany (Eckes & Althaus, 2020; for a review, see Norris & Drackert, 2018). The TestDaF speaking section assessed an examinee's ability to communicate appropriately in typical situations of university life, utilizing the Simulated Oral Proficiency Interview (SOPI) format (Kenyon, 2000). Thus, the speaking section was administered via audio-recording equipment using prerecorded prompts and printed test booklets.

The performances of 1,347 examinees on nine speaking tasks were independently scored by two out of 31 raters (26 women, 5 men) using a detailed scoring rubric. Each task was designed to target one of three increasingly higher German language proficiency levels, called TestDaF levels (*TestDaF-Niveaus*; TDN 3, TDN 4, and TDN 5). For scoring examinee performances on the three tasks targeting the highest level (TDN 5), raters used a four-category rating scale ("TDN 5 scale," for short, with categories *below TDN 3, TDN 3, TDN 4*, and *TDN 5*; coded from 0 to 3). Regarding performances on the three TDN 4 tasks, raters

Table 8. Pearson correlations between model-based rater centrality indices and raw score *SD* statistics for two different rating scales in the real-data speaking study.

	MS_W	$r_{\text{res,exp}}$	SD_τ
	Four-category rating scale		
$r_{\text{res,exp}}$	.08		
SD_τ	−.68**	−.71**	
SD_{obs}	.32	.57**	−.60**
	Three-category rating scale		
$r_{\text{res,exp}}$	−.12		
SD_τ	−.60**	−.65**	
SD_{obs}	.18	.38*	−.58**

MS_W = information-weighted mean-square fit statistic (infit) based on the FM-S rating scale model; $r_{\text{res,exp}}$ = Pearson correlation between expected scores and residuals based on the FM-S rating scale model; SD_τ = standard deviation of the Rasch-Andrich rater thresholds based on the FM-S (rater-related) partial credit model. SD_{obs} = standard deviation of observed scores.
* $p < .05$. ** $p < .01$.

used a three-category rating scale ("TDN 4 scale," with categories *below TDN 3, TDN 3*, and *TDN 4*; coded from 0 to 2). Finally, for performances on TDN 3 tasks, raters used a two-category (dichotomous) rating scale (with categories *below TDN 3* and *TDN 3*); this scale is not considered further.

Data analysis

The TestDaF speaking data had been analyzed before in a different research context (Eckes, 2005). Since the present study focused on the centrality indices' utility considering rating scales differing in length, we separately reanalyzed the TDN 5 and TDN 4 data. As in the simulation studies before, we ran the rating scale and partial credit FM-S analyses using the R package TAM (Robitzsch et al., 2020) with MMLE. For each dataset, we computed the three model-based indices (MS_W, $r_{\text{res,exp}}$, $SD(\tau_{jk})$) and the raw score *SD* index (SD_{obs}).

Results

Table 8 presents the correlations between the three model-based centrality indices and SD_{obs} computed separately for the TDN 5 and TDN 4 scales. Across scale types, a highly similar pattern of intercorrelations emerged: The infit statistic was uncorrelated with the residual-expected correlation and the SD_{obs} index, but moderately (and negatively) correlated with the SD_τ index. The other three correlations were moderately high, statistically significant, and in the expected direction.

Between-scale type correlations (i.e., same-index correlations between TDN 5 and TDN 4 index values) were insignificantly low for the infit index, $r(31) = .09$, *ns*. By contrast, the correlations for the remaining three indices were statistically significant and moderately high: $r(31) = .36$, $p < .05$ for $r_{\text{res,exp}}$, $r(31) = .57$, $p < .01$ for SD_τ, and $r(31) = .76$, $p < .01$ for SD_{obs}. Thus, specific to the infit index, the magnitude of centrality effects observed along the TDN 5 scale was not consistently related to the magnitude of centrality effects along the shorter TDN 4 scale.

Discussion

Rater centrality effects are highly resistant to change, for example, resistant to rater training targeted at reducing their impact on the assessment outcomes (e.g., Barrett, 2001; Houston & Myford, 2009; Knoch, 2011). Moreover, rater centrality, like rater severity, tends to manifest itself in a wide variety of assessment situations; that is, its occurrence does not depend on the presence of some specific situational feature or contextual cue. In this respect, centrality stands out from more context-specific rater effects, contingent on a particular context or rating situation, including halo errors, rater biases, or differential rater functioning (Myford & Wolfe, 2003; Wind & Ge, 2021; Wolfe & Song, 2016).

Given their pervasive and robust nature, it seems evident that centrality effects should be detected, measured, and accounted for in some way or other to ensure a sufficiently high level of rating quality (Eckes, 2015; Wind & Peterson, 2018). In two simulation studies, the present research evaluated three model-based statistical indices that have been proposed for centrality detection purposes (Myford & Wolfe, 2004; Wolfe & Song, 2015, 2016; Wu, 2017): weighted rater infit statistics (MS_W), residual–expected correlations ($r_{\text{res,exp}}$), and rater threshold *SD*s ($SD(\tau_{jk})$). We also examined the standard deviation of a given rater's observed score distribution (SD_{obs}), which is easy to calculate and readily available to practitioners (Johnson et al., 2009).

We studied each of these indices' sensitivity to centrality effects through comparisons with rater centrality values sampled from a latent continuum using an extended facets model – the *facets model–severity and centrality* (FM-SC; Jin & Wang, 2018). The FM-SC builds on the framework of many-facet Rasch measurement (Linacre, 1989). This framework has been adopted in various educational and psychological measurement contexts (e.g., Aryadoust et al., 2020; Eckes, 2015, 2019; Engelhard & Wind, 2018, 2021; McNamara et al., 2019). Specifically, the FM-SC extends the previous facets models by simultaneously accounting for rater severity and rater centrality. Based on the FM-SC, we sampled severity values from a normal distribution and centrality values from a lognormal distribution and held these values constant over simulation conditions.

The first simulation study addressed how well the statistical indices recovered the central tendencies. Systematically varying the number of examinees (500 or 1,000) and the number of scoring criteria (2, 3, or 4), we found, in the majority of cases, strong correlations with the true centrality measures (ω values). The second simulation study focused on the impact the range of criterion difficulties would have on the statistical indices' sensitivity to centrality. In particular, considering some researchers' observations of MS_W values greater than 1.0 for raters exhibiting central tendencies (Myford & Wolfe, 2004; Wolfe et al., 2000), we examined the correlations between the indices and ω values under conditions of low, medium, and high ranges of criterion difficulty differences. Whereas the $r_{\text{res,exp}}$, $SD(\tau_{jk})$, and SD_{obs} indices remained mostly unaffected by these variations, consistently manifesting very high degrees of sensitivity to centrality, the MS_W statistic showed a steady increase with increasing ranges. Furthermore, the higher the differences between the criterion difficulties, the more the correlations between MS_W and ω were shifted toward the positive end, becoming positive throughout when the difficulty range was high. To illustrate, under the high-range condition, a rater simulated to exhibit strong central tendencies had MS_W values around 1.5, whereas another rater simulated to exhibit extremity had MS_W values averaging below 1.0. Hence, the rater infit statistic, MS_W, does not seem suitable for detecting centrality or extremity reliably irrespective of the difficulty differences between criteria. In other words, whether central raters are associated with overfit ($MS_W < 1.0$) or misfit ($MS_W > 1.0$), depends (among other things) on the magnitude of differences between criterion difficulties. Therefore, the second simulation study provided strong support for Myford and Wolfe's (2004) admonition to be cautious about interpreting rater infit values below 1.0, for example, below the common cut-score of 0.75, indicating rater centrality.

We also reanalyzed performance ratings gathered in a large-scale speaking assessment (Eckes, 2005) to further probe the centrality indices' functioning in a real-data context. Raters used rating scales differing in length designed to fit the proficiency level targeted by each speaking task. Findings corroborated main conclusions from the simulation studies: The infit index (MS_W) correlated inconsistently with the other indices and proved uncorrelated across the two types of rating scales. The moderately strong and consistent correlations of the SD_{obs} index with the two model-based indices $r_{\text{res,exp}}$ and $SD(\tau_{jk})$ demonstrate some merit in computing the standard deviation of observed score frequency distributions for rater centrality detection. However, it should be kept in mind that using the SD_{obs} index requires a representative sample of ratings for reliably diagnosing raters' central tendencies (Wolfe & Song, 2016).

As a practical implication, detecting and measuring centrality effects can improve rater training and monitoring through providing individualized feedback to raters. This kind of feedback may help sensitize raters for possible central tendencies that otherwise go unnoticed (Houston & Myford, 2009; Knoch, 2011). Precise centrality detection is also relevant for developing efficient automated scoring systems (Shermis, 2016; Williamson et al., 2012; Wolfe, 2020). Such systems are typically calibrated using machine learning techniques based on performances previously scored by expert human raters. However, as Bridgeman (2013, p. 221) pointed out, "if the human raters are inconsistent and unreliable and/or if they overuse one or two points on the score scale, the machine cannot be effectively trained." Thus, Wind et al. (2018) found that rater effects, including centrality effects, present in the ratings used to train an automated essay scoring system were replicated in the scores this system produced. Therefore, detecting raters exhibiting central tendencies and removing their ratings from the training set of performances will likely raise the scoring system's efficiency.

Generally speaking, simulation studies are always constrained by the researchers' decision on which characteristics to include as design factors, which factor levels to consider, and which characteristics to hold constant over conditions. Within the context of performance assessments, many factors (or facets) and their mutual relations directly or indirectly influence the final scores raters assign to examinees. The conceptual–psychometric framework discussed in Eckes (2015) gives a rough idea of the complexity involved. Therefore, it should be acknowledged that simulation studies like ours are limited in generalizability to some degree.

Our focus was on the impact the examinee sample size, number of scoring criteria, and range of criterion difficulty differences had on the sensitivity of statistical centrality indices. We selected two or three levels, respectively, for each factor, and held constant characteristics of the simulated assessment setting, in particular, the rating design (incomplete, spiral), the number of raters (25), the number of rating scale categories (5), and the magnitude of the category thresholds (ranging from −0.75 to 0.75). Instead of holding these characteristics constant, they may be included as design factors to examine their impact on the indices' sensitivity in future research.

Furthermore, in operational scoring sessions, rater effects rarely come about in isolation. Instead, it is much more likely that multiple kinds of rater effects simultaneously impact on the scores assigned to examinees. Besides severity effects considered in our study, these may include halo effects, rater inaccuracy or randomness, and various forms of differential rater functioning as well as personal scoring preferences (Wang & Engelhard, 2019a, 2019b; Wind & Guo, 2019; Wolfe & Song, 2016). Therefore, another extension of our study could evaluate the performance of, and interrelations between, centrality indices under systematic variation of these combined effects' magnitude and direction.

Our simulation studies created rating data following a spiral design yielding incomplete but connected data; the real-data speaking study used a double-scoring design yielding a similar data structure (Eckes, 2015; Engelhard & Wind, 2018). Due to practical constraints, designs like these, where not every rater scores every performance, are standard in operational, large-scale assessment programs. Nonetheless, systematically varying the type of rating design, simulating, for example, designs with different proportions of missing data and comparing these with complete rating data, can provide further insight into the sensitivity of rater effect detection.

Following such a simulation approach, Stafford et al. (2018) showed that the extent of missingness did not substantially influence the performance of severity and centrality indices, that is, Rasch model-based rater location and rater threshold *SD*, respectively. Considering double-scoring designs, they found high rates of correctly detecting severe and central raters even when the proportion of examinees scored by a maximum of two randomly selected raters was as low as 5%.

Building on Stafford et al.'s simulation study, Wind and Jones (2019) focused on specific types of rating designs common in performance assessments, including systematic links, anchor performances, and spiral designs. They explored the designs' impact on the sensitivity of rater severity and centrality detection. Findings suggested that it is possible to identify rater effects regardless of the type of incomplete rating design. Wind and Ge (2021) arrived at similar conclusions regarding DRF detection in sparse assessment networks. Therefore, we expect our findings to hold under various rating designs with different proportions of missing data based on these studies.

Another deliberate decision was to construct the centrality dimension using the FM-SC (Jin & Wang, 2018). This model has demonstrated its practical utility in several large-scale language assessments (Eckes & Jin, 2021, 2021a; Jin & Wang, 2018). Considering the statistical expertise currently required to design and run an FM-SC analysis, it seems obvious that small-scale assessment situations, such as classroom-based assessments, are not the main application context. Nonetheless, detecting rater centrality effects using a direct modeling approach has unique advantages. Even though the $r_{\mathrm{res,exp}}$ and $SD(\tau_{jk})$ indices appear to do a good job under many assessment conditions, different from the FM-SC, these indices are not designed to provide examinees with scores taking simultaneously into account between-rater differences in severity and centrality. Moreover, examining rater severity and centrality estimates in conjunction may greatly help monitor raters and provide detailed feedback on their scoring behavior in operational assessment programs.

Conclusion

We have shown that the usefulness of the rater infit index (MS_W) for reliably identifying raters exhibiting central tendencies is highly questionable. Therefore, when deciding on a centrality detection statistic, we would advise researchers not to rely on rater infit statistics but instead use the model-based $r_{\mathrm{res,exp}}$ or $SD(\tau_{jk})$ indices. If the data requirements are met, using the SD_{obs} index may also be a reasonable option. If the research interest goes beyond mere centrality detection and includes estimating examinee proficiencies freed as far as possible from the impacts of rater severity and centrality, the FM-SC's direct modeling approach (Jin & Wang, 2018) has much to recommend it.

Note

1. We used the tam.mml.mfr function in TAM (Robitzsch et al., 2020), implementing marginal maximum likelihood estimation of parameters because this technique is less subject to estimation bias (von Davier, 2016; Wu et al., 2016; for a discussion of different estimation techniques, see Linacre, 2004, 2021). The TAM code for Simulation Study 1 and a simulated dataset are available on the Open Science Framework: https://osf.io/m4gbr/

ORCID

Kuan-Yu Jin http://orcid.org/0000-0002-0327-7529
Thomas Eckes http://orcid.org/0000-0002-8820-5902

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). 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

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49–58. https://openjournals.library.sydney.edu.au/index.php/IEJ/article/view/6773/7418

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 221–232). Routledge.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

DeMars, C. E. (2017). Infit and outfit: Interpreting statistical significance and magnitude of misfit in conjunction. *Journal of Applied Measurement*, *18*(2), 163–177. http://jampress.org/pubs.htm

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang. https://doi.org/10.3726/978-3-653-04844-5

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., & 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., & Jin, K.-Y. (2021a). Examining severity and centrality effects in TestDaF writing and speaking assessments: An extended Bayesian many-facet Rasch analysis. *International Journal of Testing*. Advance online publication. https://doi.org/10.1080/15305058.2021.1963260

Eckes, T., & Jin, K.-Y. (2021b). Measuring rater centrality effects in writing assessment: A Bayesian facets modeling approach. *Psychological Test and Assessment Modeling*, *63*(1), 65–94. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2021/Seiten_aus_PTAM_2021-1_ebook_4.pdf

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., & 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. (2021). A history of Rasch measurement theory. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 343–360). Routledge.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347. https://doi.org/10.1037/met0000059

Glazer, N., & Wolfe, E. W. (2020). Understanding and interpreting human scoring. *Applied Measurement in Education*, *33*(3), 191–197. https://doi.org/10.1080/08957347.2020.1750402

Houston, J. E., & Myford, C. M. (2009). Judges' perception of candidates' organization and communication, in relation to oral certification examination ratings. *Academic Medicine*, *84*(11), 1603–1609. https://doi.org/10.1097/ACM.0b013e3181bb2227

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 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

Kenyon, D. M. (2000). Tape-mediated oral proficiency testing: Considerations in developing Simulated Oral Proficiency Interviews (SOPIs). In S. Bolton (Ed.), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests* [TESTDAF: Foundations of developing a new language test] (pp. 87–106). Goethe-Institute.

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

Lane, S. (2019). Modeling rater response processes in evaluating score meaning. *Journal of Educational Measurement*, *56*(3), 653–663. https://doi.org/10.1111/jedm.12229

Lane, S., & DePascale, C. (2016). Psychometric considerations for performance-based assessments and student learning objectives. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 77–106). Routledge.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). American Council on Education/ Praeger.

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. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2004). Rasch model estimation: Further topics. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 48–72). JAM Press.

Linacre, J. M. (2021). *A user's guide to facets: Rasch-model computer programs* (Version 3.83.5). Winsteps.com. https://www.winsteps.com/manuals.htm

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

Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibrations. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 99–112). Ablex.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331–345. https://doi.org/10.1207/s15324818ame0304_3

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

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

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*(2), 159–176. https://doi.org/10.1177/014662169201600206

Myers, A. J., Ames, A. J., Leventhal, B. C., & Holzman, M. A. (2020). Validating rubric scoring processes: An application of an item response tree model. *Applied Measurement in Education*, *33*(4), 293–308. https://doi.org/10.1080/08957347.2020.1789143

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

Penfield, R. D. (2016). Fairness in test scoring. In N. J. Dorans & L. L. Cook (Eds.), *Fairness in educational assessment and measurement* (pp. 55–75). Routledge.

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard." *Applied Measurement in Education*, *28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *Package ‚TAM'* (Version 3.5-19) [Computer software]. https://cran.r-project.org/web/packages/TAM/index.html

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428. https://doi.org/10.1037/0033-2909.88.2.413

Shermis, M. D. (2016). The role of machine scoring in summative and formative assessment. In H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common core standards, smarter-balanced, PARCC, and the nationwide testing movement* (pp. 83–105). Information Age.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions*, *10*(3), 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.

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

von Davier, M. (2016). Rasch model. In W. J. Van Der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 31–48). Chapman & Hall/CRC.

Wang, J., & Engelhard, G. (2019a). Conceptualizing rater judgments and rating processes for rater-mediated assessments. *Journal of Educational Measurement*, *56*(3), 582–609. https://doi.org/10.1111/jedm.12226

Wang, J., & Engelhard, G. (2019b). 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

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

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. (2020). Exploring the impact of rater effects on person fit in rater-mediated assessments. *Educational Measurement: Issues and Practice*, *39*(4), 76–94. https://doi.org/10.1111/emip.12354

Wind, S. A., & Ge, Y. (2021). Detecting rater biases in sparse rater-mediated assessment networks. *Educational and Psychological Measurement*, *81*(5), 996–1022. https://doi.org/10.1177/0013164420988108

Wind, S. A., & Guo, W. (2019). Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and Psychological Measurement*, *79*(5), 962–987. https://doi.org/10.1177/0013164419834613

Wind, S. A., & Jones, E. (2019). The effects of incomplete rating designs in combination with rater effects. *Journal of Educational Measurement*, *56*(1), 76–100. https://doi.org/10.1111/jedm.12201

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

Wind, S. A., & Sebok-Syer, S. S. (2019). Examining differential rater functioning using a between-subgroup outfit approach. *Journal of Educational Measurement*, *56*(2), 217–250. https://doi.org/10.1111/jedm.12198

Wind, S. A., Wolfe, E. W., Engelhard, G., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing*, *18*(1), 27–49. https://doi.org/10.1080/15305058.2017.1361426

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*(1), 35–51.

Wolfe, E. W. (2020). Human scoring with automated scoring in mind. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 49–67). Routledge.

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*(2), 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.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

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

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*(4), 339–355. http://jampress.org/pubs.htm

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer. https://doi.org/10.1007/978-981-10-3302-5