

Detecting Differential Rater Functioning in Severity and Centrality: The Dual DRF Facets Model

Educational and Psychological
Measurement
2022, Vol. 82(4) 757–781
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00131644211043207
journals.sagepub.com/home/epm



Kuan-Yu Jin¹  and Thomas Eckes²

Abstract

Performance assessments heavily rely on human ratings. These ratings are typically subject to various forms of error and bias, threatening the assessment outcomes' validity and fairness. Differential rater functioning (DRF) is a special kind of threat to fairness manifesting itself in unwanted interactions between raters and performance- or construct-irrelevant factors (e.g., examinee gender, rater experience, or time of rating). Most DRF studies have focused on whether raters show differential severity toward known groups of examinees. This study expands the DRF framework and investigates the more complex case of dual DRF effects, where DRF is simultaneously present in rater severity and centrality. Adopting a facets modeling approach, we propose the dual DRF model (DDRFM) for detecting and measuring these effects. In two simulation studies, we found that dual DRF effects (a) negatively affected measurement quality and (b) can reliably be detected and compensated under the DDRFM. Using sample data from a large-scale writing assessment ($N = 1,323$), we demonstrate the practical measurement consequences of the dual DRF effects. Findings have implications for researchers and practitioners assessing the psychometric quality of ratings.

Keywords

differential rater functioning, rater bias, Bayesian estimation, MCMC

¹Hong Kong Examinations and Assessment Authority, Wan Chai, Hong Kong

²TestDaF Institute, University of Bochum, Bochum, Germany

Corresponding Author:

Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority, 7/F., Sunlight Tower, 248 Queen's Road East, Wan Chai, Kong Kong, Hong Kong.

Email: kyjin@hkeaa.edu.hk

Differential Rater Functioning

In performance assessments, it is common practice to use human raters for evaluating examinees' responses to constructed-response tasks like essay writing, providing work samples, or solving problems. However, raters may exhibit various kinds of judgmental tendencies, errors, and biases, together called rater effects, threatening the ratings' validity and fairness (Johnson et al., 2009; Myford & Wolfe, 2003; Saal et al., 1980). Therefore, researchers have developed a wide range of psychometric models and statistical indices to examine the extent to which raters are subject to these effects and to ensure sufficiently high rating quality (Engelhard & Wind, 2018; Wind & Peterson, 2018; Wolfe & Song, 2016).

A particularly intricate class of rater effects concerns differential rater functioning (DRF), commonly understood as systematic interactions between rater characteristics (e.g., response styles like severity or leniency and scoring experience) and performance- or construct-irrelevant characteristics of examinees (e.g., gender and age) or assessment conditions (e.g., time of rating and tasks or domains in an analytic scoring rubric). The net effect of such interactions is that examinee measures are not invariant over different levels of these characteristics (Eckes, 2015; Engelhard & Wind, 2018; Jin & Eckes, in press).

So far, most DRF studies have examined unwanted interactions focusing on rater severity, that is, differential severity/leniency (e.g., Hoskens & Wilson, 2001; Leckie & Baird, 2011; Lunz et al., 1996; Myford & Wolfe, 2004; Wind & Ge, 2021; Wind & Guo, 2019). Raters subject to differential severity/leniency tend to assign systematically lower/higher scores to particular subgroups of examinees after controlling for the examinees' locations on the latent variable. For example, researchers have shown that raters exhibited significant differences in severity over time (Congdon & McQueen, 2000; Lamprianou et al., 2021), examinee gender groups (Wind & Sebok-Syer, 2019), and examinee proficiency level (Kondo-Brown, 2002). Severity levels may also systematically vary when allowing raters to share their views before rating performances (Wang et al., 2014).

In the rest of this paper, we will call DRF in severity/leniency "DRF-S" (for short). Note that DRF-S is analogous to uniform differential item functioning (DIF) in two-facet data, incorporating examinees and items (Gamerman et al., 2018; Osterlind & Everson, 2009; Penfield & Camilli, 2007).

Following the Rasch facets model (RFM; Linacre, 1989), DRF-S is usually assessed through the residual-based mean square error (*MSE*) statistics infit (weighted *MSE*) and outfit (unweighted *MSE*; Engelhard, 2008; Wind & Guo, 2019; Wind & Sebok-Syer, 2019). When a rater's ratings exactly agree with RFM assumptions, this rater's infit and outfit values are expected to be close to 1.0. Wind and Guo (2019) conducted a simulation study, showing that raters exhibiting DRF-S had infit and outfit values greater than 1.0 (e.g., infit *MSE* ranging from 1.16 to 1.50). However, rater infit and outfit statistics may be influenced by many factors other than differential severity (Eckes & Jin, 2021; Wind & Guo, 2019). Therefore, observing rater fit statistics greater than 1.0 does not provide conclusive evidence of DRF-S.

Recent research has increasingly focused on another rater effect having a similarly pervasive and negative influence on rating quality: central tendency or centrality (Jin & Wang, 2018; Uto & Ueno, 2020; Wolfe & Song, 2015, 2016; Wu, 2017). This effect refers to raters' tendency to overuse the rating scale's middle category or categories. In other words, raters subject to central tendency ("central" raters, for short) underestimate high performance levels and overestimate low performance levels. In contrast to rater severity, centrality effects cannot be directly measured using a traditional many-facet-Rasch measurement framework (Eckes, 2015; Eckes & Jin, 2021).

Unlike research on DRF-S, studies investigating DRF effects in terms of differential centrality (DRF-C) have been scarce (for an exception, see Myford & Wolfe, 2009). However, similar to DRF-S, DRF-C poses threats to an assessment's validity and fairness. Therefore, both kinds of DRF effects, that is, DRF-S and DRF-C, should jointly be considered in rating quality studies.

As discussed later in somewhat more detail, DRF-C is conceptually similar to non-uniform DIF in two-facet data (i.e., examinees and items). That is, DRF-S and DRF-C may be effective in many-facet assessment settings much like uniform and non-uniform DIF in two-facet settings. We call the simultaneous presence of DRF-S and DRF-C in performance assessments "dual DRF effects." At its core, the present study proposes a new psychometric model to measure dual DRF effects. We report simulations and empirical findings attesting to the model's utility in detecting these effects.

Model Development

Rating data typically comprise at least three-facets: examinees, criteria (tasks and items), and raters. In assessment settings like this, an instance of the RFM widely used is:

$$\log \left[\frac{P_{ijkl}}{P_{ij(k-1)l}} \right] = \theta_i - \delta_j - \tau_{jk} - \eta_l, \quad (1)$$

where θ_i is the proficiency of examinee i ; δ_j and τ_{jk} are the mean difficulty and k th step difficulty of criterion j , respectively; and η_l is the severity of rater l . Higher scores will be observed when θ_i is high, δ_j is low, and η_l is low.

The distribution of step difficulties τ_{jk} is related to the variance of observed scores (Jin & Wang, 2018). For example, let there be a four-category rating scale. Figure 1 shows the item characteristic curves for two different sets of step difficulties. The dispersed steps (upper panel) would lead to a much higher proportion of middle scores than the condensed steps (lower panel).

Since raters may have different preferences for giving middle or extreme scores, it appears more appropriate to model a unique rating scale structure for each rater. Therefore, another version of the RFM to examine rater centrality looks like this (Myford & Wolfe, 2004; Wolfe & Song, 2015):

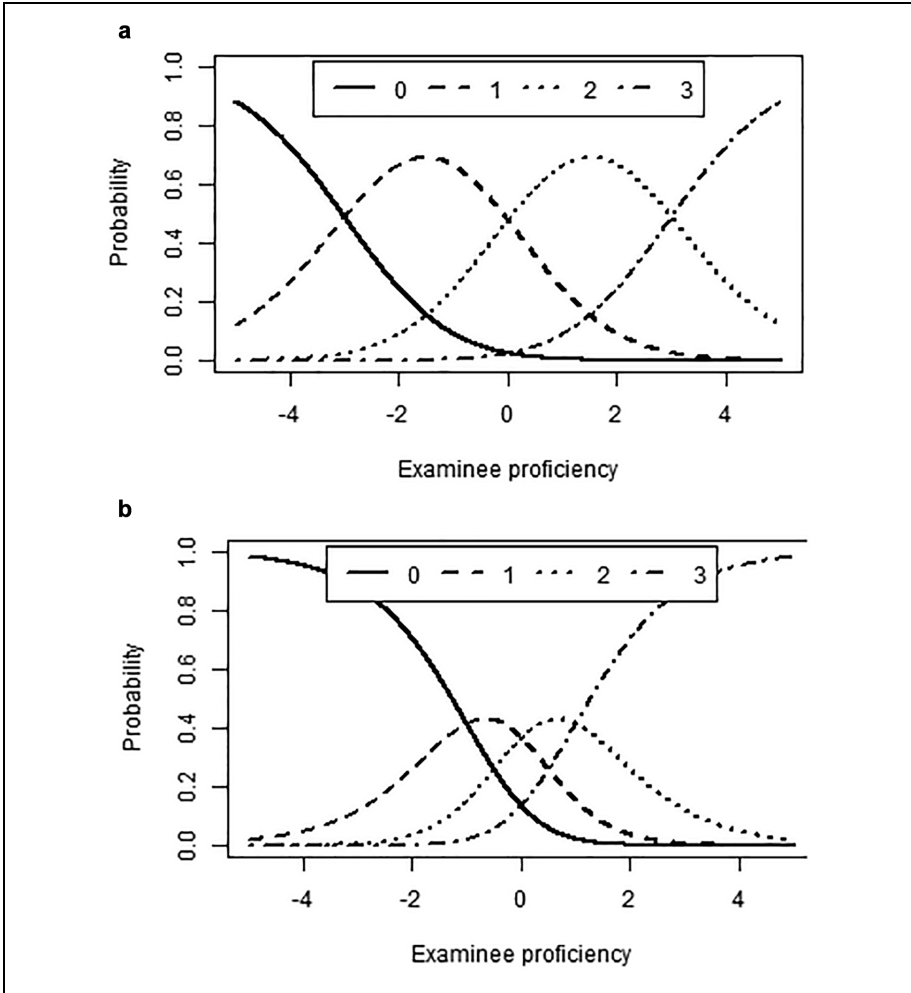


Figure 1. Two items with dispersed steps Panel (a) and condensed steps Panel (b).

$$\log \left[\frac{P_{ijkl}}{P_{ij(k-1)l}} \right] = \theta_i - \delta_j - \eta_l - v_{lk}, \tag{2}$$

where v_{lk} is the k th step difficulty for rater l . Consequently, $SD(v_{lk})$ represents the centrality level of rater l (e.g., Eckes & Jin, 2021). However, the estimate of v_{lk} is associated with measurement error; therefore, the precision of the $SD(v_{lk})$ index is problematic. Also, $SD(v_{lk})$ is not directly estimated in the model, making statistical testing infeasible.

Based on this reasoning, Jin and Wang (2018) proposed an extended model to quantify rater centrality:

$$\log \left[\frac{P_{ijkl}}{P_{ij(k-1)l}} \right] = \theta_i - \delta_j - \omega_l \tau_{jk} - \eta_l, \tag{3}$$

where ω_l is the centrality of rater l , indicating that raters are allowed to vary in the spread of individual step difficulties. In Equation (3), the parameter ω_l assumes positive values (i.e., $\omega_l > 0$). Generally, ω_j values greater than 1 indicate that raters tend to overuse the rating scale’s middle categories; ω_j values less than 1 indicate that raters tend to overuse the extreme categories.

When ω_l is re-parameterized as $e^{\varpi_l} (-\infty < \varpi_l < \infty)$, Equation (2) becomes:

$$\log \left[\frac{P_{ijkl}}{P_{ij(k-1)l}} \right] = \theta_i - \delta_j - e^{\varpi_l} \tau_{jk} - \eta_l, \tag{4}$$

Raters associated with higher ϖ parameter values tend to cluster scores around the rating scale’s middle category (or categories).

Equations (3) and (4) rest on the assumption that raters maintain a uniform level of severity and centrality over subgroups. However, some raters may be subject to DRF-S, DRF-C, or both, thus exhibiting DRF-S and DRF-C to varying degrees. Therefore, we extended Equation (4) to include DRF-S and DRF-C parameters for rater l :

$$\log \left[\frac{P_{ijklg}}{P_{ij(k-1)lg}} \right] = \theta_{ig} - \delta_j - e^{\varpi_{lg}} \tau_{jk} - \eta_{lg}, \tag{5}$$

where η_{lg} and ϖ_{lg} denote rater l ’s severity and centrality, respectively, toward examinee group g .

Put differently, $\Delta_{\eta_l} (= \eta_{lF} - \eta_{lR})$ and $\Delta_{\varpi_l} (= \varpi_{lF} - \varpi_{lR})$ represent rater l ’s DRF-S and DRF-C magnitudes regarding a reference group (R) and a focal group (F). Note that examinees’ group membership may be observed or latent (Jin & Wang, 2017). We refer to Equation (5) as the “dual DRF model” (DDRFM). When $\Delta_{\eta_l} = 0$ and $\Delta_{\varpi_l} = 0$ for all raters, suggesting raters exhibit no DRF, the DDRFM becomes Equation (4). Furthermore, when all raters are fair and exhibit the same severity and centrality level, the DDRFM reduces to the two-facet partial credit model (Masters, 1982).

Figure 2 illustrates dual DRF effects’ impact on expected scores. The area between two expected score curves for the reference and focal groups shows the direction and magnitude of DRF (Raju, 1988). Let there be a criterion with mean difficulty of 0 and three step difficulties of -1.5 , 0.5 , and 1 , respectively. In the first example (Figure 2a), rater l exhibits DRF-S only (e.g., $\eta_{lR} = 0$, $\eta_{lF} = 1$ and $\varpi_{lR} = 0$, and $\varpi_{lF} = 0$); that is, this rater consistently gives lower scores to examinees in the focal group.

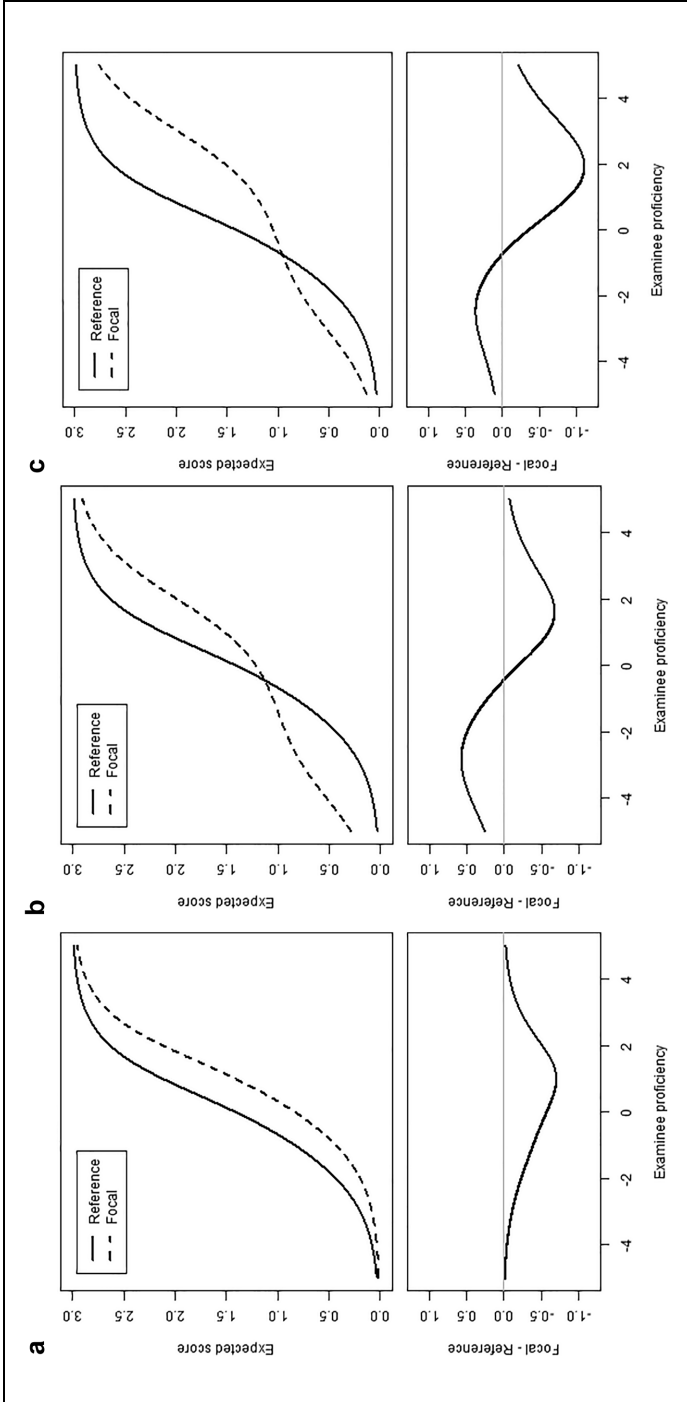


Figure 2. Influence of dual DRF effects on expected scores: (a) DRF-S, (b) DRF-C, and (c) DRF-S & DRF-C.

In the second example (Figure 2b), rater l exhibits DRF-C only (e.g., $\eta_{IR} = 0$, $\eta_{IF} = 0$, $\varpi_{IR} = 0$, and $\varpi_{IF} = 1$). Thus, focal group members with lower proficiencies would receive higher scores than reference group members at the same proficiency level; however, this pattern reverses when considering examinees at high proficiency levels.

Finally, Figure 2c displays what might happen when rater l exhibits DRF-S and DRF-C simultaneously ($\eta_{IR} = 0$, $\eta_{IF} = 1$, $\varpi_{IR} = 0$, and $\varpi_{IF} = 1$). The score differences between the focal and reference groups are more complex and difficult to interpret. Drawing the parallel between DRF and DIF effects again, the slopes and locations of the two expected score curves for the reference and focal groups would affect whether the biased rater can be successfully detected (Narayanan & Swaminathan, 1996).

In operational large-scale performance assessments, ratings are often missing by design; that is, the assessments use incomplete rating designs where not every rater scores every examinee's performance. Typically, only two or three raters score each performance, yielding proportions of missing data around 90% or more (e.g., Eckes, 2005; Wind & Jones, 2019). Therefore, we included similar conditions in our DDRFM applications.

Model Parameter Estimation

We estimated DDRFM parameters through Bayesian methods using Markov chain Monte Carlo (MCMC) in JAGS (Plummer, 2017). In the simulation and empirical studies discussed later, the priors of the estimated parameters were specified as follows: $\theta_{iR} \sim N(0, \sigma_R^2)$, $\theta_{iF} \sim N(\mu_F, \sigma_F^2)$, $\delta_i \sim N(0, 4)$, $\tau_{ik} \sim N(0, 4)$, $\eta_{IR} \sim N(0, 4)$, $\varpi_{IR} \sim N(0, 4)$, $\Delta_{\eta l} \sim N(0, 4)$, and $\Delta_{\varpi l} \sim N(0, 4)$. The priors for the hyperparameters were: $\mu_F \sim N(0, 4)$, $\sigma_R^2 \sim \lambda(0.25, 0.25)$, and $\sigma_F^2 \sim \lambda(0.25, 0.25)$. Finally, the posterior distributions of model parameters were proportional to the likelihood of the rating data and the given priors:

$$\begin{aligned}
 &g(\boldsymbol{\theta}_R, \boldsymbol{\theta}_F, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\eta}_R, \boldsymbol{\varpi}_R, \boldsymbol{\Delta}_\eta, \boldsymbol{\Delta}_\varpi | \mathbf{Y}) \\
 &\propto L(\mathbf{Y} | \boldsymbol{\theta}_R, \boldsymbol{\theta}_F, \boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\eta}_R, \boldsymbol{\varpi}_R, \boldsymbol{\Delta}_\eta, \boldsymbol{\Delta}_\varpi) \\
 &\times g(\boldsymbol{\delta}, \boldsymbol{\tau}, \boldsymbol{\eta}_R, \boldsymbol{\varpi}_R, \boldsymbol{\Delta}_\eta, \boldsymbol{\Delta}_\varpi) \times g(\boldsymbol{\theta}_R, \boldsymbol{\theta}_F | \sigma_R^2, \mu_F, \sigma_F^2) \times g(\sigma_R^2, \mu_F, \sigma_F^2),
 \end{aligned} \tag{6}$$

where \mathbf{Y} refers to the rating data and $g(\cdot)$ denotes the probability density function.

An anchor has to be set in DRF detection studies for model identification. One may constraint $\sum_{l=1}^L \Delta_{\eta l} = 0$ and $\sum_{l=1}^L \Delta_{\varpi l} = 0$, implying that the magnitude of DRF-S and DRF-C effects, respectively, is zero on average. In our studies, we opted for this constraint because researchers, before the analysis, usually lack the knowledge to identify raters that may be considered fair or unbiased. Following this way of anchoring, each rater received an individual measure for DRF-S and DRF-C. Alternatively,

when a rater (indexed as L') has been chosen to act as a reference (an unbiased or expert rater), this rater may serve as an anchor using a modified constraint, for example, by setting the DRF-S and DRF-C for rater L' to zero: $\Delta_{\eta L'} = 0$ and $\Delta_{\omega L'} = 0$.

Across studies we ran a single MCMC chain to save computing time.¹ We discarded the first 5,000 iterations as burn-in, and kept the subsequent 5,000 iterations to form the posterior distributions. Based on the final 5,000 draws, we used the posterior distributions' means as point estimates for the respective parameters. For a sample of the JAGS code, see the Supplemental Appendix.

We examined the convergence of MCMC draws within a chain by computing the Geweke z statistic (test of non-stationarity; Geweke, 1992; see also Jackman, 2009). The distributions of the first and second halves of samples for each estimate were compared. When converged, the two distributions are not significantly different, and the Geweke z statistic would follow an asymptotically standard normal distribution. In other words, when the Geweke z statistic exceeds ± 1.96 (i.e., the 95% confidence interval), the convergence of MCMC is considered questionable. We expected the percentage of z statistics exceeding ± 1.96 to be around 5% for a converged model.

The Bayesian chi-square statistic with posterior predictive model checking (Rubin, 1984) was used to evaluate the DDRFM's absolute goodness-of-fit. The posterior predictive p -value (PPP-value) summarizes the discrepancy between the observed and replicated responses given the parameter estimates in each iteration. An extreme PPP-value (higher than .975 or lower than .025) indicates poor data-model fit (Levy & Mislevy, 2016).

Finally, to address the issue of relative model fit, that is, to compare the DDRFM to the RFM in terms of data-model fit, we computed the Bayesian deviance information criterion (DIC; Spiegelhalter et al., 2002) for each model. Models showing smaller DIC values are generally preferred as better fitting (Levy & Mislevy, 2016).

Simulation 1: Consequences of Ignoring Dual DRF

Design

Simulation 1 aimed to examine the consequences of ignoring dual DRF effects. We created rating data for 200 examinees, five criteria, and three raters; these conditions are quite common in applied assessment research (e.g., Kondo-Brown, 2002; Springer & Bradley, 2018). Data generation followed the same general settings as in Jin and Wang (2018). Ratings were provided using a five-category rating scale (ranging from 0 to 4).

Table 1 shows the rating design underlying this simulation. The total sample was divided equally into a reference group (R) and a focal group (F). Within each group, there were two equally sized subgroups. Within each subgroup, two raters assigned ratings to each examinee. We defined the first rater as unbiased (fair) and the second rater as biased (subject to severity or centrality).

Specifically, for examinees belonging to Subgroups 1 and 2, the first rater, rater U (for short), was an unbiased rater, defined by $\eta_{UR} = \eta_{UF} = 0$ and $\omega_{UR} = \omega_{UF} = 0$.

Table 1. Rating Design Used in Simulation Study 1.

Group	Subgroup	n	Raters		
			Rater U	Rater S	Rater C
Reference	1	50	X	X	
Focal	2	50	X	X	
Reference	3	50	X		X
Focal	4	50	X		X

Note. Rater U is an unbiased (fair) rater. Rater S is a biased rater exhibiting DRF-S. Rater C is a biased rater exhibiting DRF-C.

The second rater, rater S, was biased, defined by $\eta_{SR} = 0$, $\eta_{SF} = 1$ and $\varpi_{SR} = \varpi_{SF} = 0$; that is, rater S was simulated to be severe when rating focal group examinees but not when rating reference group examinees, thus exhibiting DRF-S.

For examinees belonging to Subgroups 3 and 4, the first rater was again unbiased and the second rater again biased. However, this time, the biased rater, rater C, was defined by $\eta_{CR} = \eta_{CF} = 0$ and $\varpi_{CR} = 0$, $\varpi_{CF} = 1$; that is, rater C was simulated to assign scores clustering around the rating scale’s middle category when rating focal group examinees but not when rating reference group examinees, thus exhibiting DRF-C.

For each subgroup, we generated 100 datasets from the DDRFM. Each dataset was fit by the RFM and the DDRFM. In these analyses, the parameters for rater U were fixed at their true values, whereas the parameters for raters S and C were freely estimated. In each replication, we computed the assessment’s reliability as the squared correlation coefficient between the true and estimated examinee proficiencies (θ values) under the RFM and DDRFM. We also computed the mean absolute rank change (*MARC*) to evaluate the empirical consequences of using these two-facet models:

$$MARC = \frac{\sum_{i=1}^{200} |\zeta_i - \hat{\zeta}_i|}{200}, \tag{7}$$

where ζ_i and $\hat{\zeta}_i$ are the true and estimated rank orders of examinee i in each replication. We hypothesized that ignoring dual DRF effects by fitting the RFM would decrease test reliability and increase *MARC*.

Results

The Markov chains converged in all analyses. Unsurprisingly, the DIC uniformly favored the DDRFM (i.e., the data-generating model) across replications. Figure 3 displays the distributions of scores observed within each subgroup across 100 replications. The fair and biased raters assigned almost identical scores to reference group

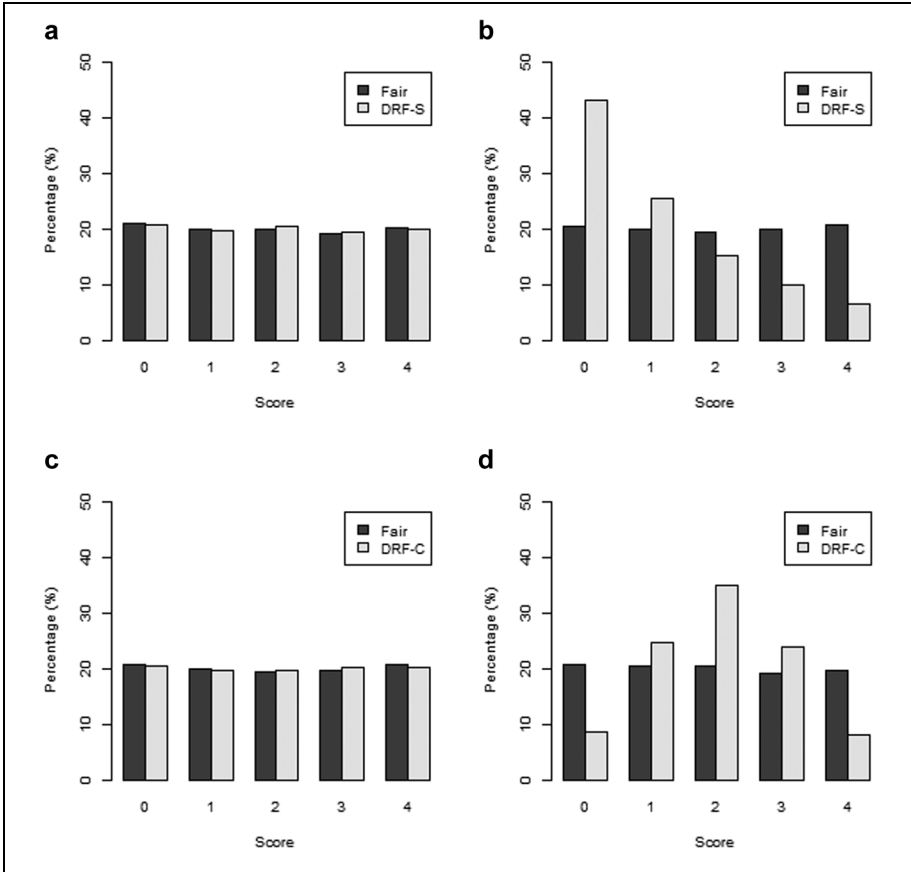


Figure 3. Observed score distributions (percentages) in Simulation Study I: (a) Subgroup 1, (b) Subgroup 2, (c) Subgroup 3, and (d) Subgroup 4.

examinees (Subgroups 1 and 3; Figure 3a and c). However, for focal group examinees (Subgroups 2 and 4), the distributions of scores assigned by fair and biased raters were strikingly different. In Subgroup 2 (Figure 3b), biased rater S generally assigned lower scores ($M = 2.12$) to examinees in the focal group than fair rater U ($M = 3.00$). In Subgroup 4 (Figure 3d), biased rater C tended to assign much more scores around the middle categories ($SD = 1.07$) than fair rater U ($SD = 1.41$). In sum, the two different kinds of biased ratings each had a strong differential impact on individual examinees' scores belonging to the focal group.

Figure 4 displays test reliability and *MARC* distributions for the reference and focal groups under the two models across 100 replications. Reliability values were consistently higher under the DDRFM than under the RFM (upper panel).

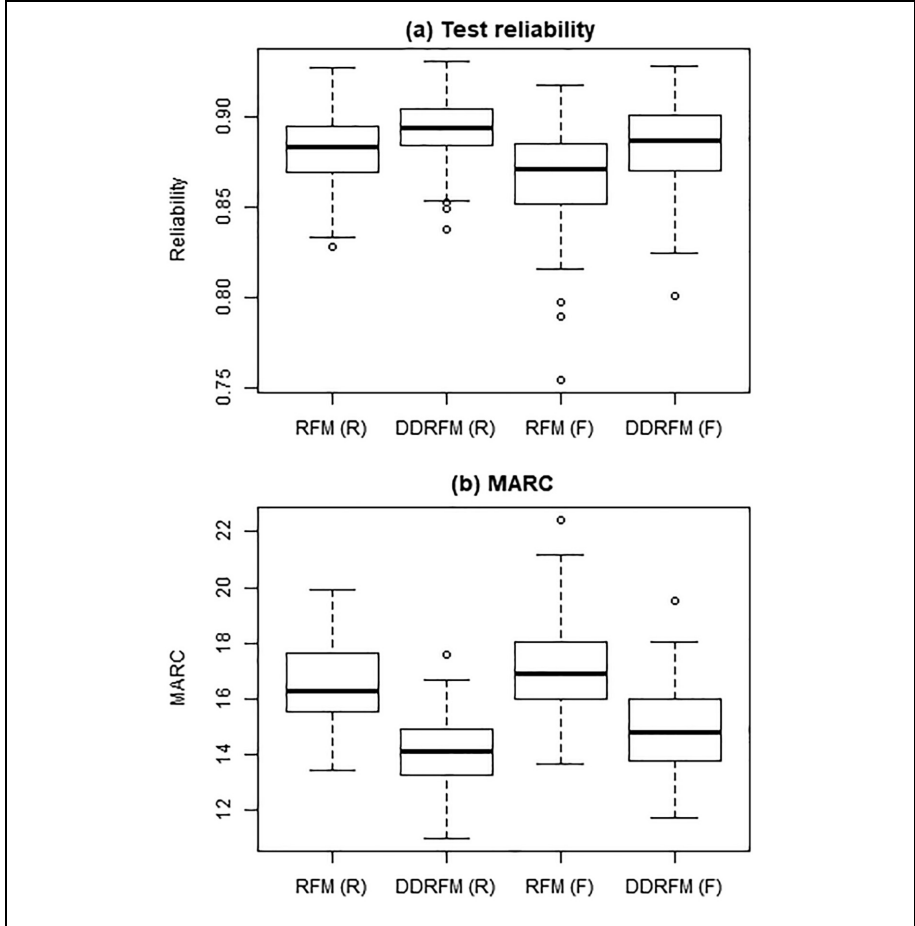


Figure 4. Test reliability estimates and mean absolute rank change (*MARC*) for reference (R) and focal group examinees (F) under the RFM and the DDRFM in Simulation Study I.

Furthermore, *MARC* values were consistently lower under the DDRFM than under the RFM (lower panel).

Regarding the biased ratings' impact on the measurement precision, we looked at the relationship between true and estimated examinee proficiency estimates under the RFM and the DDRFM. For an exemplary simulated dataset, Figure 5 illustrates this relationship. Under the RFM (upper panel), the values were scattered widely around the identity line. By contrast, under the DDRFM, the values generally stayed much closer to the identity line, providing further evidence of higher DDRFM measurement precision.

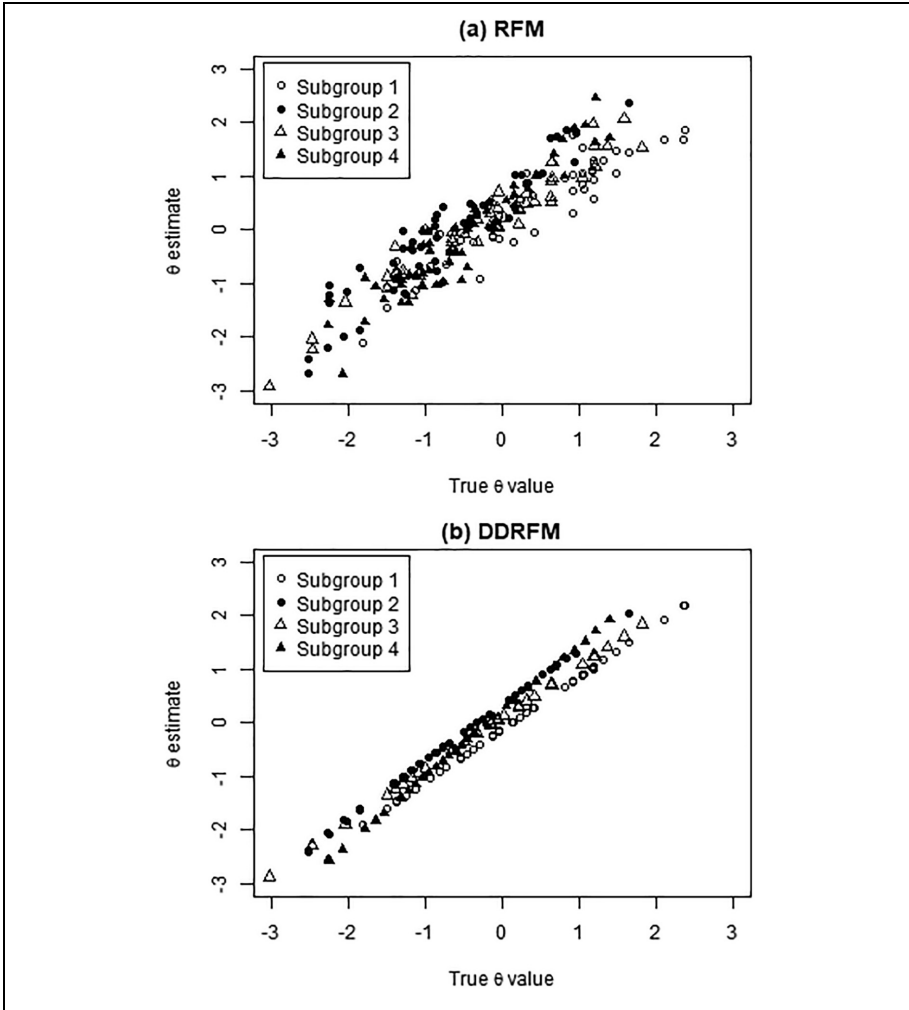


Figure 5. Relationship between true and estimated examinee proficiency estimates under the RFM and the DDRFM in Simulation Study 1.

Simulation 2: False-positive and True-positive Rates of Dual DRF Detection

Design

In Simulation Study 2, we examined the efficiency of the model-based approach to detect dual DRF effects correctly. We generated item responses under the DDRFM employing a design similar to Simulation Study 1. Differences concerned the

Table 2. Type I Error Rates (%) of Severity and Centrality Effect Detection under the DDRFM in Simulation Study 2.

Rater	True value	DRF size	Three criteria		Five criteria	
			Three scale categories	Five scale categories	Three scale categories	Five scale categories
Severity						
1 ^a	0	0	—	—	—	—
2	0	0.5	—	—	—	—
3	0	0	8/5	6/7	7/6	12/7
4	0	0.5	—	—	—	—
5	-0.5	0	6/13	4/4	5/9	6/8
6	-0.5	0	6/5	7/4	3/9	12/5
7	-0.5	0	7/6	6/5	4/7	10/9
8	0.5	0	4/5	10/2	5/5	8/9
9	0.5	0	5/8	7/4	2/6	4/12
10	0.5	0	1/4	4/4	7/6	7/6
Centrality						
1 ^a	0	0	—	—	—	—
2	-0.5	0	1/3	3/6	4/2	6/4
3	0	0	10/9	4/8	6/4	4/5
4	0.5	0	9/1	6/6	7/6	3/2
5	-0.5	0	0/4	5/5	2/4	4/6
6	0	0.5	—	—	—	—
7	0.5	0	8/6	3/6	4/3	6/3
8	-0.5	0	1/7	5/6	1/4	7/5
9	0	0.5	—	—	—	—
10	0.5	0	4/9	6/7	5/4	7/3

Note. Values before and after the slash are Type I error rates under the conditions with 500 and 1,000 examinees, respectively.

^aRater 1 is an unbiased rater used as an anchor. Type I error rates are reported for raters who did not exhibit DRF-S or DRF-C.

systematic variation of three factors: (1) the number of examinees ($I = 500$ or $1,000$), (2) the number of criteria ($J = 3$ or 5), and (3) the number of rating scale categories ($K = 3$ or 5). These three factors are directly related to the amount of information on how raters use the rating scale.

Reference and focal groups were sampled from a standard normal distribution with a mean of zero and unit variance. The true values of parameters were taken from Jin and Wang (2017) and related studies. When $J = 3$, the mean criterion difficulties were set at $-0.5, 0,$ and 0.5 ; when $J = 5$, the mean difficulties were set at $-0.5, -0.25, 0, 0.25,$ and 0.5 . For simplicity, the step parameters for all criteria were set at -0.5 and 0.5 when $K = 3$, and $-0.75, -0.25, 0.25,$ and 0.75 when $K = 5$.

As shown in Tables 2 and 3, we specified three rater severity levels (η), each with 3 or 4 raters: $-0.5, 0,$ and 0.5 . Similarly, we specified three rater centrality levels (ω): $-0.5, 0,$ and 0.5 . The size of DRF-S (Δ_{η}) was set at 0.5 for the two biased raters,

Table 3. Power Rates (%) of Severity and Centrality Effect Detection under the DDRFM in Simulation Study 2.

Rater	True value	DRF size	Three criteria		Five criteria	
			Three scale categories	Five scale categories	Three scale categories	Five scale categories
Severity						
1 ^a	0	0	—	—	—	—
2	0	0.5	41/72	57/94	65/87	89/100
3	0	0	—	—	—	—
4	0	0.5	30/62	41/74	44/74	62/83
5	-0.5	0	—	—	—	—
6	-0.5	0	—	—	—	—
7	-0.5	0	—	—	—	—
8	0.5	0	—	—	—	—
9	0.5	0	—	—	—	—
10	0.5	0	—	—	—	—
Centrality						
1 ^a	0	0	—	—	—	—
2	-0.5	0	—	—	—	—
3	0	0	—	—	—	—
4	0.5	0	—	—	—	—
5	-0.5	0	—	—	—	—
6	0	0.5	26/49	47/77	25/57	69/96
7	0.5	0	—	—	—	—
8	-0.5	0	—	—	—	—
9	0	0.5	19/52	49/73	28/65	73/93
10	0.5	0	—	—	—	—

Note. Values before and after the slash are power rates under the conditions with 500 and 1,000 examinees, respectively.

^aRater 1 is an unbiased rater used as an anchor. Power rates are reported for raters who did exhibit DRF-S or DRF-C.

and the size of DRF-C (Δ_{mf}) was set at 0.5 for the other two biased raters; that is, two raters were more severe toward the focal group, whereas the other two raters exhibited more centrality toward the focal group.

The simulated data were fit with the data-generating model. Therefore, rater *l* was identified as exhibiting DRF-S (or DRF-C) if the 95% probability interval of Δ_{η_l} (or Δ_{mf}) did not include zero. Regarding the Type I error rate, we computed the percentage of times a fair rater was mistakenly identified as exhibiting DRF-S (or DRF-C) across 100 replications. As to the power rate, we computed the percentage of times a biased rater was correctly identified as exhibiting DRF-S (or DRF-C). Under the true model, the Type I error rates should be close to the nominal 5% level across conditions. The power rates should depend on the manipulated factors: We expected higher power rates under conditions with more examinees, more criteria, and more scale categories.

Results

Table 2 summarizes the Type I error rates under the eight conditions. Overall, Type I error rates of DRF-S and DRF-C detection were well controlled at the 5% level across conditions. Similarly, Table 3 presents the power rates under the eight conditions. Consistently, higher power rates were observed with 1,000 examinees, 5 criteria, and 5 scale categories.

We also conducted three-way ANOVAs on the power rates of DRF-S and DRF-C, respectively. Under the current simulation design, the first manipulated factor (i.e., the number of examinees) had the greatest influence on the power rate of DRF-S detection ($p = .002$, partial $\eta^2 = 0.70$). In contrast, the third manipulated factor (i.e., the number of rating scale categories) had the greatest influence on the power rate of DRF-C detection ($p < .001$, partial $\eta^2 = 0.98$).

Notably, rater centrality (ω parameter) influenced the power rates of DRF-S detection: a biased rater with lower centrality was more likely detected as exhibiting DRF-S than a biased rater with higher centrality. The reason for this centrality-dependent DRF-S detection is mainly that, for a given DRF-S size, a larger area between two expected score curves for the reference and focal groups is obtained for raters with lower ω parameters.

An Empirical Example: TestDaF Essay Rating Data

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 writing section assesses an examinee's ability to produce a coherent and well-structured text on a given topic taken from the academic context. The dataset considered here had been analyzed before using a traditional facets modeling approach (Eckes, 2005).

Two out of 29 raters (23 women and 6 men) independently scored the written performances of 1,359 examinees on three criteria (global impression, task fulfillment, and linguistic realization) using a four-category rating scale (*below TDN 3*, *TDN 3*, *TDN 4*, and *TDN 5*; coded as 0–3). As mentioned previously, the proportion of missing ratings was high (93.1%). Nonetheless, the dataset was connected (Eckes, 2015; Engelhard & Wind, 2018).

At the time of this exam (April 2002), the examinees' full names were provided in the label attached to each paper. Raters were thus able to infer most examinees' gender from parts of these names. Moreover, research has shown that female and male handwritings look significantly different (Beech & Mackintosh, 2005; Boulet & McKinley, 2005; Siddiqi et al., 2015), increasing the chances for raters to guess an examinee's gender correctly. Therefore, the essay ratings seemed to provide a suitable dataset for studying the potential impact of gender-related differential

Table 4. Means and Standard Deviations of Observed Scores Assigned to Female and Male Examinees by 29 Raters in the Essay Rating Study.

Rater	Female examinees			Male examinees		
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>
1	46	1.69	0.89	35	1.56	0.96
2	65	2.10	0.82	45	1.96	0.89
3	88	1.63	0.61	60	1.43	0.76
4	30	1.98	0.79	18	2.06	0.68
5	73	1.40	1.04	45	1.13	1.06
6	16	1.08	0.82	10	1.63	0.76
7*	44	1.75	0.89	13	2.00	0.95
8	54	2.35	0.78	33	2.28	0.70
9	46	1.67	0.86	48	1.32	0.87
10	31	2.00	0.88	23	1.93	0.75
11	61	1.56	0.63	63	1.48	0.71
12	21	1.52	1.06	14	1.19	0.89
13	55	1.44	0.90	54	1.66	0.85
14	104	2.08	0.77	76	1.50	0.85
15	54	1.42	0.95	51	1.42	0.89
16	80	2.01	0.76	52	1.68	0.69
17	50	1.73	0.85	47	1.35	0.76
18	69	1.71	0.97	73	1.47	0.93
19	44	1.67	1.04	52	1.38	1.10
20	122	2.30	0.82	100	2.06	0.90
21*	39	2.09	1.11	46	1.09	1.04
22	18	2.28	0.79	24	2.07	0.89
23*	54	0.94	0.95	40	1.02	0.93
24*	20	1.97	0.69	17	1.86	0.78
25	19	1.30	0.98	16	1.23	0.69
26	18	1.93	0.80	27	1.72	0.76
27	19	1.88	0.73	13	2.10	0.64
28*	41	1.66	1.01	54	1.60	1.03
29*	75	1.64	1.06	41	1.44	1.04

Note. Raters marked with an asterisk (*) are male. *N* is the number of female and male examinees, respectively, each rater scored. *M* and *SD* values refer to the four-category rating scale ranging from 0 (below *TDN* 3) to 3 (*TDN* 5). Higher means indicate higher writing proficiency.

severity and centrality effects. Since 36 examinees did not indicate their gender, we included the responses of 1,323 examinees (728 females and 595 males) in the present analyses.

Table 4 gives the means and standard deviations (*SDs*) of each rater's score distribution, listed separately for female and male examinees. Observed means and *SDs* can serve as rough indicators of rater severity and centrality, respectively (Johnson et al., 2009). Across raters, the overall mean for female examinees was 1.78 (*SD* = 0.84). For male examinees, the overall mean was 1.58 (*SD* = 0.83), suggesting that females, on average, slightly outperformed males. Notably, however, 6 raters

(four women and two men) on average assigned higher ratings to male examinees, pointing to the possibility that raters' severity levels may have varied with gender. Similarly, the observed *SDs* show that 16 raters (13 women and 3 men) tended to assign females less homogeneous scores than males. Our DDRFM analyses were to shed more light on these differential tendencies.

Data Analysis

The RFM and DDRFM were fit to the data, with male examinees treated as the reference group. For model identification, the mean ability of males was fixed at zero. Following the partial-credit approach in Eckes (2005), each criterion was modeled to have its own set of threshold parameters.

Results

The PPP-value of the Bayesian chi-square for the DDRFM was .175, indicating satisfactory absolute data-model fit. Furthermore, only 6.6% of the *z* statistics fell beyond the ± 1.96 interval, suggesting convergence of the Markov chain to the posterior distribution. The DDRFM yielded a lower DIC value (13,315) than the RFM (13,612), providing evidence that the more complex model (i.e., DDRFM) fit the essay rating data better, taking into account the greater number of estimated parameters in terms of the penalty statistic for the DDRFM.

The findings regarding model fit suggest that important assumptions of the RFM regarding rater functioning, including equal centrality levels across raters and the nonexistence of DRF-S or DRF-C, respectively, were disconfirmed. On the other hand, the DDRFM fit the rating data so well that a closer look at the resulting parameter estimates seemed warranted.

Figure 6 displays the estimates (and the 95% probability intervals) for DRF-S and DRF-C, respectively. As can be seen (Figure 6a), raters 8, 13, 15, and 23 (three women and one man) were more severe to females, whereas raters 11, 14, 19, and 29 (three women and one man) were more lenient to females. The DRF-C estimates (Figure 6b) reveal that (female) rater 3 and (male) rater 7 tended to overuse middle categories (*TDN* 3 or *TDN* 4) when rating females' performances, whereas (female) raters 8 and 16 tended to preferably assign extreme categories (*below TDN* 3 or *TDN* 5) to females. Notably, rater 8 was identified to exhibit DRF-S and DRF-C simultaneously.

Under the DDRFM, for male examinees, the mean θ estimate was 0 (males were treated as the reference group), and the variance of the estimates was 6.09; for female examinees, the respective statistics were 0.61 and 6.60, respectively, suggesting that females outperformed males. Figure 7a displays the relationship between the θ estimates obtained under the DDRFM and the RFM. The two sets of estimates were highly correlated (0.99). However, when taking the DDRFM θ estimates as the gold standard, the rank-order change was between -89 and 136 ranks ($M = 8.85$) for males and between -175 and 216 ranks ($M = 7.23$) for females. As illustrated in

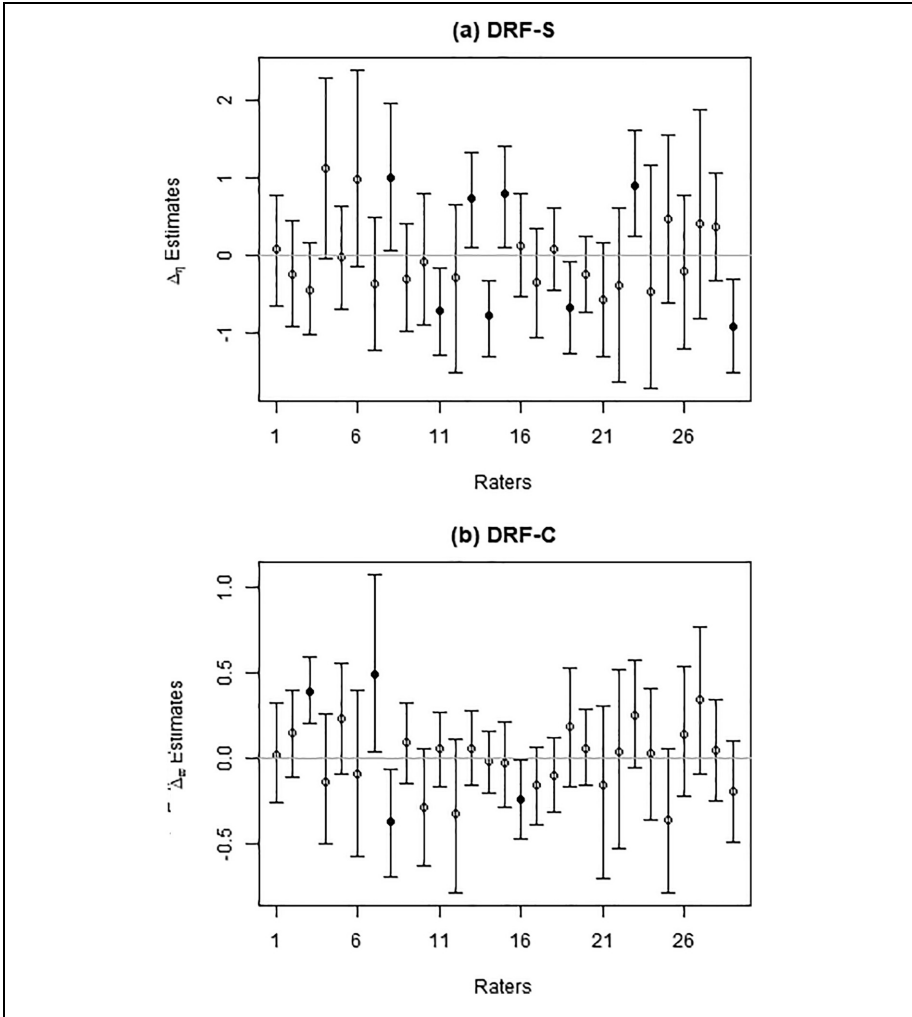


Figure 6. DRF-S and DRF-C estimates for 29 raters in the TestDaF essay rating study.
Note. Hollow and solid circles refer to non-DRF and DRF raters, respectively.

Figure 7b and c, fitting the RFM would make many male examinees attain higher ranking positions than female examinees, suggesting that females are disadvantaged when dual DRF effects are not taken into account.

Discussion

Severity and centrality are two main kinds of rater characteristics that need to be detected, measured, and compensated for as much as possible to ensure performance

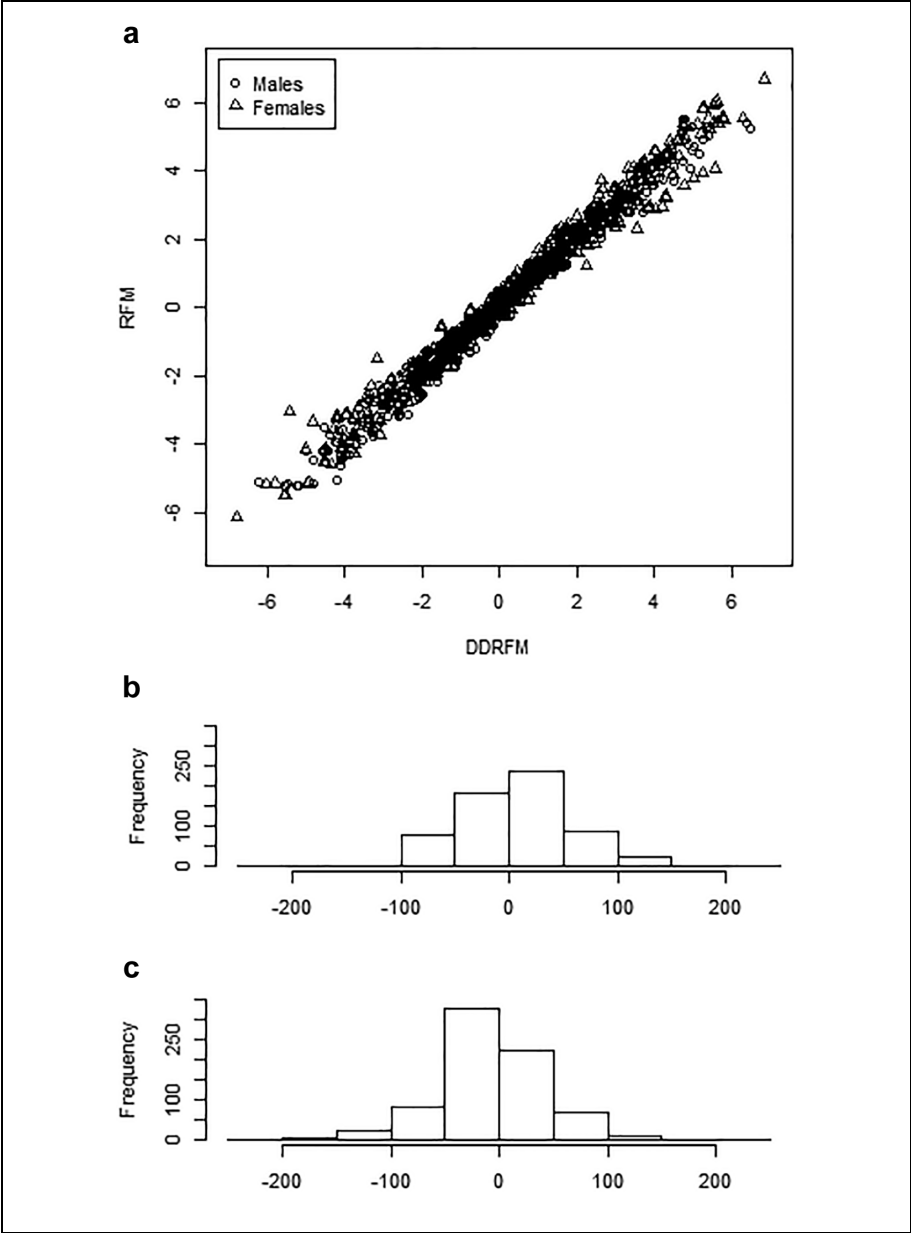


Figure 7. Writing proficiency estimates under the RFM and the DDRFM in the essay rating study: (a) θ estimates, (b) Rank-order change for males, and (c) Rank-order change for females.

assessments' validity and fairness (Engelhard, 1992; Myford & Wolfe, 2003, 2004; Saal et al., 1980). Ideally, once trained and sufficiently experienced, raters should exhibit only minimal severity and centrality levels, if at all, when judging examinee performances. Consistently small severity and centrality effects imply that raters should assign ratings, which are largely invariant over examinee characteristics irrelevant to the performance or construct in question. However, the bulk of research on rating quality in a wide variety of assessment settings has consistently provided evidence that raters, even when extensively trained, are subject to various forms of errors and biases, including DRF (Eckes, 2015; Engelhard & Wind, 2018; McNamara et al., 2019; Wind & Peterson, 2018; Wolfe & Song, 2016). In other words, performance assessments typically do not seem to come close to the ideal of raters acting as a homogeneous group and achieving undisputable high levels of interrater agreement and reliability. Instead, it may reasonably be doubted whether this ideal is ever attainable by rater training alone.

Measurement models help remedy this problematic situation. In the present research, we proposed a facets model, the dual differential rater functioning model (DDRFM), aiming to examine whether raters exhibited differential severity (DRF-S), differential centrality (DRF-C), or both, toward identifiable groups of examinees. In two simulation studies, we found that ignoring DRF-S and DRF-C would lead to poorer measurement quality, especially for the focal group. In addition, the proposed DDRFM allows researchers to detect DRF-S and DRF-C with well-controlled Type I error rates.

We also applied the DDRFM to a real dataset, using ratings from a large-scale writing performance assessment. The presence of gender-related severity biases in these data had been studied before using an exploratory interaction analysis based on the RFM (Eckes, 2005). In pairwise comparisons, statistically significant results were obtained for only three raters: two raters were more severe with male than with female examinees, one rater was more lenient with male than with female examinees. The present DDRFM analysis yielded a much more detailed, precise, and comprehensive picture of gender bias: 11 out of the 29 raters were identified to exhibit differential severity or differential centrality regarding examinee gender groups. One of these raters was subject to both DRF-S and DRF-C.

We illustrated the practical implications of dual DRF effects by comparing the examinee rank-ordering resulting from the DDRFM proficiency estimates (reference) to the examinee rank-ordering produced by the RFM estimates. On average, the rank orderings for male examinees differed by 8.85 ranks; for female examinees, the rank-order change averaged 7.23, depending on which model was used for estimating their proficiency. Rank differences of this magnitude may have severe consequences for individual examinees, for example, when deciding on university admission.

Several limitations should also be noted. In this study, we focused on between-rater variations of DRF-S and DRF-C. More specifically, in the DDRFM (Equation 5), we treated the dual DRF effects as fixed effects. Consequently, we were not able to consider possible within-rater variations (Wang & Wilson, 2005). A straightforward approach to account for within-rater differences in dual DRF effects would be

to extend the DDRFM by defining the rater parameters as random effects (e.g., following normal distributions). Of course, fitting a random-effects DDRFM to real data requires much larger sample sizes for parameter estimation; that is, each rater should have rated a sufficiently large number of performances.

It should be noted that the choice of priors may influence the estimation results for DDRFM parameters (Gelman et al., 2013 ; Lunn et al., 2013). Therefore, it appears generally advisable to reanalyze the data with different priors and check the consistency of the results. In Simulation Study 1, for example, we ran an additional analysis with even less informative priors to investigate the parameter recovery when fitting the RFM and DDRFM; that is, we used $N(0, 10)$ and $\lambda(0.01, 0.01)$, respectively. The results were highly similar to the first analysis, attesting to our findings' stability.

Furthermore, non-rating data targeting the same latent proficiency may be included as auxiliary information to improve parameter estimation. Particularly, it is common to use multiple item or task formats in large-scale assessments, including selected-response items (e.g., multiple-choice or short-answer questions; Guo & Wind, 2021; Wind & Ge, 2021). In such mixed-format situations, examinee responses to items objectively scored could be included to provide more information on examinee proficiency, further increasing the precision of dual DRF effects detection in DDRFM studies.

As a practical implication, measuring both differential severity and differential centrality effects can improve rater training and monitoring through providing individualized feedback to raters. This kind of feedback may help sensitize raters for possible biases that otherwise would go unnoticed. Regarding test or assessment development, any clues that may reveal examinees' performance- or construct-irrelevant characteristics should be eliminated as far as possible. As for the TestDaF writing assessment (Eckes, 2005), subsequent examinations used completely anonymized paper scanning and rating procedures. Furthermore, with the recent advent of the web-based, digital TestDaF (g.a.s.t, 2020), examinees type their written responses using a keyboard, eliminating any gender-related information that may emerge from their handwriting.

Much like the situation in DIF research more generally (Penfield & Camilli, 2007), providing reasonable explanations for dual DRF effects that a DDRFM analysis may have revealed is quite challenging and often requires considering information on examinees and raters, respectively, coming from other sources. For example, since rating accuracy is influenced by raters' cognitive and meta-analysis strategies (Zhang, 2016), researchers may investigate through a mix of quantitative and qualitative methods (e.g., eye-tracking analysis, think-aloud protocols, or structured interviews) how raters subject to dual DRF effects go about assigning ratings to examinees. Findings from studies along these lines may help to identify possible causes of dual DRF effects.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kuan-Yu Jin  <https://orcid.org/0000-0002-0327-7529>

Supplemental Material

Supplemental material for this article is available online.

Note

1. In our simulations, we found that, for the RFM and DDRFM, a single chain yielded results just as stable as multiple chains.

References

- Beech, J. R., & Mackintosh, I. C. (2005). Do differences in sex hormones affect handwriting style? Evidence from digit ratio and sex role identity as determinants of the sex of handwriting. *Personality and Individual Differences, 39*(2), 459–468. <https://doi.org/10.1016/j.paid.2005.01.024>
- Boulet, J. R., & McKinley, D. W. (2005). Investigating gender-related construct-irrelevant components of scores on the written assessment exercise of a high-stakes certification assessment. *Advances in Health Sciences Education, 10*(1), 53–63. <https://doi.org/10.1007/s10459-004-4297-y>
- Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163–178. <https://doi.org/10.1111/j.1745-3984.2000.tb01081.x>
- 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. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang.
- 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.
- Eckes, T., & Jin, K.-Y. (2021). 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. (2008). Differential rater functioning. *Rasch Measurement Transactions, 21*(3), 1124. <https://www.rasch.org/rmt/rmt213f.htm>

- Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.
- Gamerman, D., Gonçalves, F. B., & Soares, T. M. (2018). Differential item functioning. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 3, pp. 67–86). Chapman & Hall/CRC.
- g.a.s.t. (2020). Der digitale TestDaF: Zielsetzung, Konzept und Testformat [The digital TestDaF: objective, conceptualization, and test design]. Gesellschaft für Akademische Studienvorbereitung und Testentwicklung (g.a.s.t. e.V.). <https://www.testdaf.de/fileadmin/testdaf/downloads/Broschueren/Der-digitale-TestDaF.pdf>
- 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.
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, & J. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 169–193). Oxford University Press.
- Guo, W., & Wind, S. A. (2021). Examining the impacts of ignoring rater effects in mixed-format tests. *Journal of Educational Measurement*. <https://doi.org/10.1111/jedm.12292>
- Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State examination. *Journal of Educational Measurement*, 38(2), 121–145. <https://doi.org/10.1111/j.1745-3984.2001.tb01119.x>
- Jackman, S. (2009). *Bayesian analysis for the social sciences*. Wiley.
- Jin, K.-Y., & Eckes, T. (in press). Detecting rater centrality effects in performance assessments: A model-based comparison of centrality indices. *Measurement: Interdisciplinary Research and Perspectives*. <https://doi.org/10.1080/15366367.2021.1972654>
- Jin, K.-Y., & Wang, W.-C. (2017). Assessment of differential rater functioning in latent classes with new mixture facets models. *Multivariate Behavioral Research*, 52(3), 391–402. <https://doi.org/10.1080/00273171.2017.1299615>
- 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 Press.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, 19(1), 3–31. <https://doi.org/10.1191/0265532202lt218oa>
- Lamprianou, I., Tsagari, D., & Kyriakou, N. (2021). The longitudinal stability of rating characteristics in an EFL examination: Methodological and substantive considerations. *Language Testing*, 38(2), 273–301. <https://doi.org/10.1177/0265532220940960>
- Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399–418. <https://doi.org/10.1111/j.1745-3984.2011.00152.x>
- Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Chapman & Hall/CRC.
- Linacre, J. M., (1989). *Many-facet Rasch measurement*. MESA Press.
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian analysis*. Chapman & Hall/CRC.
- Lunz, E. M., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibration. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 99–112). Ablex.

- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/bf02296272>
- McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: The role of measurement*. Oxford University Press.
- Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386–422.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189–227.
- Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, 46(4), 371–389. <https://doi.org/10.1111/j.1745-3984.2009.00088.x>
- Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, 20(3), 257–274. <https://doi.org/10.1177/014662169602000306>
- Norris, J., & Drackert, A. (2018). Test review: TestDaF. *Language Testing*, 35(1), 149–157. <https://doi.org/10.1177/0265532217715848>
- Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). SAGE Publications.
- Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay, & C. R. Rao (Eds.), *Handbook of statistics* (Vol. 26, pp. 125–167). Elsevier.
- Plummer, M. (2017). JAGS version 4.3.0 user manual. https://people.stat.sc.edu/hansont/stat740/jags_user_manual.pdf
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53(4), 495–502. <https://doi.org/10.1007/bf02294403>
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12(4), 1151–1172. <https://doi.org/10.1214/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(2), 413–428. <https://doi.org/10.1037/0033-2909.88.2.413>
- Siddiqi, I., Djeddi, C., Raza, A., & Souici-meslati, L. (2015). Automatic analysis of handwriting for gender classification. *Pattern Analysis and Applications*, 18(4), 887–899. <https://doi.org/10.1007/s10044-014-0371-0>
- 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(4), 583–639. <https://doi.org/10.1111/1467-9868.00353>
- Springer, D. G., & Bradley, K. D. (2018). Investigating adjudicator bias in concert band evaluations: An application of the many-facets Rasch model. *Musicae Scientiae*, 22(3), 377–393. <https://doi.org/10.1177/1029864917697782>
- 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., Su, C.-M., & Qiu, X.-L. (2014). Item response models for local dependence among multiple ratings. *Journal of Educational Measurement*, 51(3), 260–280. <https://doi.org/10.1111/jedm.12045>

- 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>
- 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>
- Wolfe, E. W., & Song, T. (2015). Comparison of models and indices for detecting rater centrality. *Journal of Applied Measurement, 16*(3), 228–241.
- 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.
- 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
- Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing, 27*(1), 37–53. <https://doi.org/10.1016/j.asw.2015.11.001>