Psychological Test and Assessment Modeling, Volume 64, 2022 (1), 87-111

Detecting Illusory Halo Effects in Rater-Mediated Assessment: A Mixture Rasch Facets Modeling Approach

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

Abstract

Halo effects refer to a common source of error in human judgment. In rater-mediated assessments where each rater assigns multiple scores to examinees, raters subject to halo tend to give similar scores on conceptually distinct traits, dimensions, or criteria. An intricate problem with empirically detecting halo effects concerns the separation between illusory halo due to judgmental biases or cognitive distortions, and true halo, due to actual overlap between the traits or criteria used for scoring examinee performances. The present research used the mixture Rasch facets model for halo effects (MRFM-H; Jin & Chiu, 2022) to detect illusory halo. In two separate studies, raters scored examinees' writing performances on a set of criteria using a four-category rating scale. Halo parameters were estimated building on Bayesian Markov chain Monte Carlo methods implemented in the freeware JAGS run within the R environment. The findings revealed that (a) the MRFM-H fit the data well but not better than the basic Rasch facets model (RFM), (b) in Study 2, we identified three raters that may have been subject to illusory halo effects. The discussion focuses on practical implications for ensuring high rating quality in performance assessments.

Keywords: rater effects, illusory halo, mixture Rasch model, Bayesian statistics, model fit

[1] *Correspondence concerning this article should be addressed to:* Thomas Eckes, PhD, TestDaF Institute, University of Bochum, Universitätsstr. 134, 44799 Bochum, Germany; e-mail: thomas.eckes@gast.de, https://orcid.org/0000-0002-8820-5902

[2] Hong Kong Examinations and Assessment Authority, Hong Kong

Halo effects, or halo errors, have long been a concern for researchers and practitioners involving human raters to evaluate examinee responses or performances (Cooper, 1981; Guilford, 1936; Thorndike, 1920). Typically, halo effects are an issue in rating designs where each rater assigns multiple scores to the same examinee, as when raters use an analytic scoring rubric to rate a given performance on distinct traits, dimensions, or criteria (Balzer & Sulsky, 1992; Cooper, 1981). For example, raters may assign a separate score to content, organization, and language use when evaluating writing performances. In assessment contexts like this, raters subject to halo tend to give similar scores on conceptually distinct criteria (Myford & Wolfe, 2004).

As commonly acknowledged, halo effects can originate from different cognitive or judgmental processes. Fisicaro and Lance (1990) distinguished three kinds of processes: First, a particular rater may form a general impression of an examinee's performance, which directly impacts the evaluation on each criterion (e.g., a first impression based on physical appearance, age, ethnic, or gender stereotypes); second, the rater's perception of an examinee's performance may be dominated by a single, subjectively salient performance feature (e.g., legible handwriting, strong voice); third, the rater may fail to discriminate adequately between the intended meanings of the criteria (e.g., due to lack of familiarity with scoring guidelines, fatigue). Whatever processes may be involved in any given instance, the final result is a biased evaluation of an examinee's performance, threatening the validity and fairness of the assessment outcomes. Put differently, the amount of unique information conveyed by each single criterion score is significantly reduced in the presence of halo effects.

Therefore, detecting halo effects plays an essential role in ensuring high rating quality in performance assessments (Knoch et al., 2021; Wind & Peterson, 2018; Wolfe & Song, 2016). Over the years and decades, many methods, statistics, and measures have been proposed for that purpose (Balzer & Sulsky, 1992; Cooper, 1981; Saal et al., 1980). More common statistics include the mean intercorrelation among scores on different criteria or the within-examinee variability (standard deviation) across criteria (Fisicaro & Vance, 1994; Murphy, 1982). Following these statistics' rationale, halo effects would manifest in increased mean criterion intercorrelations or lowered within-examinee standard deviations.

More sophisticated approaches rest on measurement models, particularly the many-facet Rasch measurement or facets modeling framework (Linacre, 1989). For example, researchers discussed several methods and statistics for detecting halo effects building on facets models (Engelhard, 2002; McNamara & Adams, 2000; Myford & Wolfe, 2004). However, most of these statistics, such as residual-fit statistics, often fail to provide conclusive evidence of halo effects. Thus, Myford and Wolfe (2004) cautioned that "interpretation of the fit indices is not straightforward, but rather context bound" (p. 211). A relevant context factor is the variation of criterion or trait difficulties (Myford & Wolfe, 2004): In an assessment where criterion difficulties show little variation, halo effects would be indicated by mean-square fit statistics less than 1 (overfit); conversely, when criterion difficulties show large variation, halo effects would be indicated by mean-square fit statistics greater than 1 (misfit). To

complicate matters, the extent to which criterion difficulties vary is usually unknown before analyzing the rating data and may differ from rater to rater (Marais & Andrich, 2011).

There is a considerable body of research showing that facets models are well suited to measure and compensate for well-known rater effects, including severity/leniency (Eckes, 2005, 2015; Engelhard, 2002; Engelhard & Wind, 2018; McNamara et al., 2019; Myford & Wolfe, 2003, 2004) and, more recently, centrality/extremity (Eckes & Jin, 2021a, 2021b; Jin & Eckes, in press; Jin & Wang, 2018) as well as differential rater functioning (Jin & Eckes, 2021; Jin & Wang, 2017). However, facets models specifically addressing the detection and measurement of halo effects have long been lacking. Jin and Chiu (2022) advanced a facets modeling approach that closes this gap. In the present research, we apply the Jin and Chiu model to two real data sets drawn from rater-mediated writing assessments and discuss the model's implications for studying halo effects in rater-mediated assessments more generally.

True and Illusory Halo Effects

Previous attempts at identifying halo effects have faced a specific challenge: Many criteria, traits, or dimensions used in analytic scoring are correlated with one another simply because they relate to the same construct, latent variable, or performance the assessment is targeting. This kind of intrinsic relation among traits or criteria gives rise to what has been called "true" or "valid" halo, as opposed to "illusory" or "invalid" halo (e.g., Bartlett, 1983; Murphy, 1982; Pulakos et al., 1986). Murphy et al. (1993) expressed this distinction clearly:

> "Except in those rare circumstances in which the dimensions being rated are truly orthogonal, the observed correlation between dimensional ratings represents a composite of the true correlation and the net result of the cognitive distortions, errors in observation and judgment, and rating tendencies of the individual rater (i.e., illusory halo)" (p. 220).

Several researchers proposed methods to distinguish true from illusory halo (Balzer & Sulsky, 1992; Bechger et al., 2010; Lai et al., 2015). Most of these methods require using a special kind of rating design. For example, Balzer and Sulsky (1992) suggested collecting expert ratings, assumed to be largely unaffected by halo, and comparing these "true" ratings with the ratings provided by operational raters. According to this approach, illusory halo would be indicated, for example, by trait intercorrelations for operational ratings closer to 1.0 than the corresponding value for expert ratings. However, expert ratings are not readily available in many applied settings, incur extra costs, and may themselves be subject to illusory halo to some extent.

Lai et al. (2015) suggested an even more demanding rating design. Their proposal entails comparing scores obtained when a single rater evaluates examinee performances on all traits or criteria (single-rater multi-trait design) to scores obtained when different raters each evaluate examinee performances on only a single trait (multi-rater single-trait design; for a similar approach, see Bechger et al., 2010). Lai et al. provided evidence that ratings from a multi-rater single-trait design may be less subject to illusory halo than ratings from a single-rater multi-trait design. However, as the authors admit, rating designs involving multiple raters (each rating a different trait) are more difficult and expensive to implement than single-rater designs. They also called attention to the fact that illusory halo can still exist in single-trait designs. For example, raters may be influenced similarly by the same performance features, constituting what Marais and Andrich (2011) called a common rater halo effect. In retrospect, these caveats seem to lend substance to Murphy et al.'s (1993) earlier conclusion that "it is impossible to separate true from illusory halo in most field settings" (p. 223).

However, as discussed in the next section, this conclusion is no longer valid in its generality. Jin and Chiu's (2022) new facets modeling approach estimates the extent to which individual raters are subject to illusory halo, thus empirically separating true from illusory halo. This separation is accomplished building on rating designs commonly used in rater-mediated assessment settings.

The Mixture Rasch Facets Model for Halo Effects (Jin & Chiu, 2022)

Facets models in widespread use today rest on Linacre's (1989) many-facet extensions of the Rasch rating scale model (RSM; Andrich, 1978) or partial credit model (PCM; Masters, 1982). As a starting point for developing a facets model capable of detecting illusory halo, Jin and Chiu (2022) considered a typical three-facet assessment situation where J raters assign scores to N examinees on I criteria using a rating scale with $m + 1$ categories, that is, $k = 0, \ldots, m$. When employing a PCM instantiation, the Rasch facets model (RFM) may be specified as follows:

$$\ln\left[\frac{p_{nikj}}{p_{ni(k-1)j}}\right] = \theta_n - (\beta_i + \tau_{ik}) - \alpha_j, \qquad (1)$$

where p_{nikj} is the probability of examinee n receiving on criterion i a rating of k from rater j, $p_{ni(k-1)j}$ is the probability of examinee n receiving on criterion i a rating of $k - 1$ from rater j, θ_n is the ability of examinee n, β_i is the difficulty of criterion i, τ_{ik} is the difficulty of receiving on criterion i a rating of k relative to $k - 1$, and α_j is the severity of rater j.

In Equation 1, the parameter τ_{ik} is the Rasch–Andrich threshold parameter (Andrich, 1998; Linacre, 2006) defined for a particular criterion. It is assumed that the rating

scale structure varies from criterion to criterion. In other words, the model shown in Equation 1 is a three-facet criterion-related PCM accounting for rater severity. Alternatively, an RSM version could be specified by defining the threshold parameter as constant across criteria (i.e., replacing, in Eq. 1, τ_{ik} by τ_k). Such a model would impose the same set of threshold values for all raters and criteria. From a substantive point of view, the RSM version of the RFM implies that raters use the rating scale categories across criteria in the same manner (Eckes, 2015; Linacre, 2000). We adopted the less restrictive approach and estimated criterion-specific threshold values in the present research. Each criterion is allowed to have a unique rating scale structure. Therefore, we examined how the rating scale functioned for different criteria.

Jin and Chiu (2022) extended the RFM given above to capture the distinction between illusory and true halo effects. The *mixture Rasch facets model for halo effects* (MRFM-H) includes a latent rater severity dimension and two latent classes for raters, that is, the class of "normal" (or "no-halo") raters and the class of "(illusory) halo raters". Specifically, the MRFM-H is given as follows:

$$\ln\left[\frac{p_{nikj}}{p_{ni(k-1)j}}\right] = \theta_n - (1 - x_j)(\beta_i + \tau_{ik}) - x_j(\beta^* + \tau_k^*) - \alpha_j, \qquad (2)$$

where all parameters are as in Equation 1 except for the x_j parameter and the β^* and the τ_k^* terms discussed below.

The x_j parameter is a binary variable following a Bernoulli distribution π; this variable indicates the latent class membership of rater *j*. That is, if rater *j* exhibits illusory halo, $x_j = 1$; otherwise, $x_j = 0$. Hence, the likelihood of the observed ratings reveals the latent class to which rater *j* belongs. When $x_j = 0$ for each rater $j \in J$, all raters are normal raters showing no illusory halo effect. In this case, the MRFM-H reduces to the RFM (Jin & Chiu, 2022).

The β^* and the τ_k^* terms designate the difficulty and the threshold values, respectively, of a *generalized criterion* i^*:

$$\beta^* = \beta_1 = \ldots = \beta_I, \qquad (3)$$

$$\tau_k^* = \tau_{1k} = \ldots = \tau_{Ik}. \qquad (4)$$

The generalized criterion i^* represents a combination of the individual, analytic criteria when raters are subject to halo effects and, therefore, do not distinguish between criteria, leading them to assign each examinee similar scores across the criteria (Myford & Wolfe, 2004). In this case, the criteria will not differ significantly in terms of their difficulty and the associated threshold values, respectively. For model identification, the constraint $\Sigma\beta_i = 0$ is required; therefore, β^* equals zero, implying that the generalized criterion represents halo raters' overall evaluation standard.

Regarding the MRFM-H (Eq. 2), criteria possessing homogeneous psychometric characteristics (i.e., highly similar β and τ parameters, respectively) yield similar ratings, indicating true halo effects. Hence, the conditional probability of a normal rater assigning a score will be identical across criteria. Conversely, when criteria possess

heterogeneous psychometric characteristics (i.e., when β and τ parameters, respectively, vary greatly across criteria), similar ratings are probably caused by illusory halo effects. In this case, the conditional likelihoods of a normal rating will differ substantially from those of an illusory halo rating. Therefore, the x_j parameter in Equation 2 represents the magnitude of an illusory halo effect for rater j.

In a simulation study, Jin and Chiu (2022) demonstrated the efficiency of the MRFM-H to detect illusory halo. Building on a Bayesian estimation approach, they showed that applying the MRFM-H yielded over 99% recovery of raters' normal vs. illusory halo class membership (raters were classified as illusory halo raters when the x_j estimate exceeded .5). When this model was fitted to data where raters exhibited no halo effects (i.e., when the RFM was the true model), the MRFM-H recovered parameter estimation (rater severity, criterion difficulty, and thresholds) as well as the RFM. Applications of the MRFM-H to several real datasets yielded further evidence of the model's efficiency.

The present research applied the MRFM-H (Eq. 2) to real datasets from a different assessment context, thereby widening the scope of this model's use. Also, we conducted a more detailed analysis of the MRFM-H data–model fit, compared the MRFM-H parameter estimates to those provided by the RFM, and looked at the model's practical implications. The datasets were the same as previously analyzed using facets models designed to detect rater centrality effects (Eckes & Jin, 2021). Therefore, besides using datasets possessing well-known characteristics, this reanalysis had the additional advantage of comparing parameter estimates across different models and analyses, lending substance to conclusions drawn from the analyses. Also, the Study 1 dataset is publicly available (Eckes, 2019; Eckes & Jin, 2021).[3]

Specifically, both datasets reanalyzed here were from rater-mediated writing assessments administered in the context of admission to higher education institutions in Germany. In two separate language proficiency examinations (called Study 1 and Study 2, hereafter), raters scored examinees' writing performances on a set of criteria using a four-category rating scale. Differences between the two studies lay primarily in (a) the way raters were assigned to examinee performances and (b) the kind and number of criteria included in the scoring rubric (more detail on these differences is provided later).

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

Research questions

We reanalyzed two datasets from writing assessments using Jin and Chiu's (2022) mixture Rasch facets model for halo effects (MRFM-H). Based on the MRFM-H and adopting Bayesian parameter estimation procedures, we investigated the extent to which raters were subject to illusory halo. We also ran the Rasch facets model (RFM) on the same data to compare with the MRFM-H findings. Following this methodological approach, the present research aimed to answer the following three questions:

1. How does the MRFM-H fit the writing assessment data?
2. How does the MRFM-H compare to the RFM in terms of data–model fit?
3. Does the MRFM-H identify individual raters exhibiting illusory halo effects?

Method

Participants

In both studies, the examinees were international students applying for entry to higher education institutions in Germany. Raters were specialists in German as a foreign language trained and monitored to comply with the scoring guidelines. In Study 1, 18 raters scored the writing performances of 307 examinees; in Study 2, 12 raters scored the writing performances of 206 examinees (see Table 1 for more information on examinee and rater samples).

Table 1: Study 1 and Study 2 assessment design features

Feature	Study 1	Study 2
Facets		
Examinees (*N*)	307	206
Females (%)	158 (51.5)	140 (68.0)
Males (%)	149 (48.5)	66 (32.0)
Raters (*J*)	18	12
Females (%)	14 (77.8)	11 (91.7)
Males (%)	4 (22.2)	1 (8.3)
Criteria (*I*)	3	9
Rating scale categories	4	4
Number of essays rated		
Min	19	29
Max	68	30
M	36.0	29.9
Proportion of missing ratings (%)	88.3	85.5

Note. In Study 1 and Study 2, the rating design was incomplete but connected (performance links design).

Instruments and procedure

In each study, the writing task was part of the Test of German as a Foreign Language (TestDaF, *Test Deutsch als Fremdsprache*) – an officially recognized language exam for international students applying for entry to higher education institutions in Germany (Eckes & Althaus, 2020). Examinee performance in each of four test sections (reading, listening, writing, and speaking) is related to one of three increasingly higher levels of language proficiency, the TestDaF levels (*TestDaF-Niveaus*, TDNs). For a review of the TestDaF, see Norris and Drackert (2018).

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

The Study 1 and Study 2 rating designs were incomplete but connected (Eckes, 2015). In other words, the designs corresponded to the performance links design Jin and Chiu (2022) used in their simulation study. Specifically, in Study 1, two raters rated each performance independently on the writing task (i.e., each essay). Also, one rater provided ratings of two randomly selected essays from each of the other 17 raters' workload. Raters scored the essays using a four-category scale, the TDN scale, with categories *below TDN 3*, *TDN 3*, *TDN 4*, and *TDN 5*. For computations, *below TDN 3* was scored "2", and the other levels were scored from "3" to "5".

Ratings were provided separately on three criteria: *global impression* (referring to lower-level aspects such as fluency, train of thought, and structure), *task fulfillment* (completeness, description, and argumentation), and *linguistic realization* (breadth of syntactic elements, vocabulary, and correctness). Following the Study 1 rating design, there were 648 ratings; that is, 614 double ratings plus 34 third ratings, making a total of 1,944 ratings (the proportion of missing ratings was 88.3%).

The Study 2 rating data were collected as part of a larger validation program, focusing on the TestDaF writing, speaking, and listening sections. A total of 206 examinee performances were sampled from the entire set of 3,949 essays produced by TestDaF examinees in April 2012 (2,557 females, 1,392 males). Performances covered the whole range of TDN levels. Following the Study 2 rating design, all 12 raters independently scored the same subset of 10 randomly selected essays; most of the remaining essays were each rated by a single rater (some of these essays were also rated by two raters each to strengthen the link between raters).

Different from Study 1, raters in Study 2 scored the essays separately on each of the lower-level aspects *fluency*, *train of thought*, and *structure* (replacing the higher-level *global impression* criterion); *completeness*, *description*, and *argumentation* (replacing the *task fulfillment* criterion); and *breadth of syntactic elements*, *vocabulary*, and *correctness* (replacing the *linguistic realization* criterion). One rater inadvertently returned scores for only 29 examinees; the other 11 raters provided scores for 30 examinees, resulting in a set of 359 ratings on each of the nine criteria. Thus, a total of 3,231 ratings was available for estimating RFM and MRFM-H parameters (the proportion of missing ratings was 85.5%). Table 1 summarizes the design features characterizing Study 1 and Study 2.

Data analysis

We estimated the model parameters building on a Bayesian approach (Gelman et al., 2013; Lunn et al., 2013). In particular, Bayesian estimation was performed using Markov chain Monte Carlo (MCMC) techniques implemented in the JAGS freeware (JAGS = Just Another Gibbs Sampler; Plummer, 2017). The R2jags package (Su & Yajima, 2021) was employed to run the MCMC models in JAGS. This package provides interface functions to facilitate running user-specified MCMC models within R (R Core Team, 2021).

Following Jin and Chiu (2022), we specified the prior distributions of the model parameters as follows:

$$\theta_n \sim N(\mu, 1/\sigma^2), \quad (5)$$

$$\beta_i \sim N(0, 0.1), \quad (6)$$

$$\alpha_j \sim N(0, 0.1), \quad (7)$$

$$\tau_{ik} \sim N(0, 0.1), \quad (8)$$

$$x_j \sim Bernoulli(\pi), \quad (9)$$

where $N(\mu, \tau)$ is the normal distribution with mean μ and precision τ, for $\tau > 0$; the variance σ^2 of the normal distribution is $1/\tau$; *Bernoulli*(p) is the Bernoulli distribution with probability p (Lunn et al., 2013).

Also, we used the following priors for the hyperparameters:

$$\mu \sim N(0, 0.1), \quad (10)$$

$$1/\sigma^2 \sim Gamma(0.1, 0.1), \quad (11)$$

$$\pi \sim Beta(1, 1), \quad (12)$$

where *Gamma*(r, λ) is the Gamma distribution with shape r and rate λ; *Beta*(α, β) is the Beta distribution with shape parameters α and β (Lunn et al., 2013).

Three MCMC chains were run to assess convergence to the posterior distribution. The initial 5,000 draws were discarded in each chain as burn-in, and the draws from the subsequent 5,000 iterations were retained for parameter estimation. The mean of the posterior distributions was used as the point estimate, or expected a-posteriori (EAP) estimate, of a given parameter; similarly, the posterior standard deviation was used as an estimate of the standard (or model) error associated with a parameter estimate. The gap between posterior draws was set at 10 to reduce the autocorrelation effect; that is, every 11th posterior draw was recorded (Levy & Mislevy, 2016). These specifications were the same in Study 1 and Study 2.

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

Using the posterior predictive model-checking (PPMC) method, we examined the fit of the (observed) data to the model (Gelman et al., 2013; Levy & Mislevy, 2016). This method compares the observed data with the data generated or predicted by the model. In particular, the PPMC approach involves computing a discrepancy measure using each simulated value from the posterior distributions for the parameters. Plotting the distribution of these values (the realized values) against the posterior predictive

values' distribution provides a graphical display of data–model fit, which may be summarized in the tail-area probability, also known as the posterior predictive *p*-value (PPP-value). Extreme PPP-values (i.e., values close to 0 or 1) indicate poor data–model fit; medium values, that is, values around .5, indicate much better fit (Levy & Mislevy, 2016, p. 242).

Finally, to address the issue of relative model fit, three different criteria were computed. The first criterion was the Bayesian deviance information criterion (DIC; Spiegelhalter et al., 2002; van der Linde, 2005). Models showing smaller DIC values generally fit better (Levy & Mislevy, 2016, p. 248).

Some research has indicated that DIC may not function well when considering more complex models, particularly models including latent classes (Gelman et al., 2013; Merkle et al., 2019; Spiegelhalter et al., 2014). Therefore, we computed two other Bayesian model comparison methods: the Watanabe–Akaike Information Criterion (WAIC; Watanabe, 2010) and the Leave-One-Out Information Criterion (LOOIC; Geisser & Eddy, 1979), both available in the R package loo (Vehtari et al., 2020). Unlike DIC, WAIC and LOOIC require the use of the whole posterior distribution instead of point estimates, which is why these two criteria are viewed as fully Bayesian (Gelman et al., 2013, 2014; Luo & Al-Harbi, 2017).

Results

Data–model fit

Tables 2 and 3 summarize the convergence and data–model fit statistics. In both studies and for each parameter under the RFM and the MRFM-H, respectively, the potential scale reduction factor (PSRF) values were close to 1.0, indicating that the MCMC chains converged to the target (posterior) distribution without problems. Also, the PPP-values were non-extreme or close to .5, confirming that, in each instance, the data–model fit was satisfactorily high.

The DIC statistics for Study 1 and Study 2 show that the RFM fit the data slightly better than the MRFM-H. However, differences of the magnitude observed here (i.e., less than 5) are commonly not considered as favoring one model over the other (Gelman et al., 2013; Lunn et al., 2013). Much the same conclusions hold for the WAIC and LOOIC statistics; the differences are negligibly small in size. In other words, for the present data sets, there do not seem to be any substantial differences in RFM and the MRFM-H regarding model fit.

Table 2: Bayesian model fit and comparison statistics, Study 1

Statistic	RFM	MRFM-H
PSRF (min–max)		
Examinee proficiency	1.000–1.025	1.000–1.021
Rater severity	1.003–1.018	1.003–1.016
Criterion difficulty	1.001–1.002	1.000–1.007
Thresholds	1.000–1.003	1.001–1.009
PPP-value	.500	.487
DIC	3,307.6	3,308.4
WAIC	3,218.1	3,220.4
LOOIC	3,232.4	3,235.4

Note. RFM = Rasch facets model. MRFM-H = mixture Rasch facets model for halo effects. Throughout the RFM and MRFM-H analyses, the partial credit versions were used. PSRF = Proportional scale reduction factor. PPP-value = Posterior predictive *p*-value. DIC = Deviance information criterion. WAIC = Watanabe–Akaike information criterion. LOOIC = Leave-one-out information criterion.

Table 3: Bayesian model fit and comparison statistics, Study 2

Statistic	RFM	MRFM-H
PSRF (min–max)		
Examinee proficiency	1.000–1.025	1.000–1.009
Rater severity	1.007–1.042	1.002–1.008
Criterion difficulty	1.000–1.002	1.000–1.007
Thresholds	1.000–1.005	1.000–1.007
PPP-value	.331	.327
DIC	5,869.6	5,881.7
WAIC	5,842.4	5,843.3
LOOIC	5,847.0	5,848.2

Note. RFM = Rasch facets model. MRFM-H = mixture Rasch facets model for halo effects. Throughout the RFM and MRFM-H analyses, the partial credit versions were used. PSRF = Proportional scale reduction factor. PPP-value = Posterior predictive *p*-value. DIC = Deviance information criterion. WAIC = Watanabe–Akaike information criterion. LOOIC = Leave-one-out information criterion.

Criterion difficulty and threshold parameter estimates

Table 4 presents the RFM and the MRFM-H estimates of criterion difficulty and criterion-specific thresholds (Study 1). With both kinds of analysis, *global impression* was much less difficult than *task fulfillment* and *linguistic realization*, respectively; also, these two criteria proved to be similarly difficult – a finding in line with earlier analyses of the same data using different facets models (Eckes, 2015; Eckes & Jin, 2021). Similarly, the close correspondence of threshold estimates across criteria indicates that the raters used and interpreted the rating scale in much the same way irrespective of the criterion considered.

Table 4: Bayesian criterion difficulty and threshold estimates, Study 1

	RFM		MRFM-H	
Criterion	Estimate	*SE*	Estimate	*SE*
Global impression				
β_1	–0.77	0.07	–0.83	0.09
τ_1	–2.78	0.19	–2.78	0.19
τ_2	–0.24	0.12	–0.25	0.14
τ_3	3.02	0.16	3.03	0.16
Task fulfillment				
β_2	0.37	0.06	0.40	0.07
τ_1	–2.87	0.17	–2.89	0.17
τ_2	–0.22	0.12	–0.20	0.12
τ_3	3.09	0.16	3.10	0.16
Linguistic realization				
β_3	0.40	0.07	0.43	0.08
τ_1	–3.12	0.17	–3.14	0.17
τ_2	0.02	0.12	0.00	0.12
τ_3	3.11	0.17	3.14	0.17

Note. RFM = Rasch facets model. MRFM-H = mixture Rasch facets model for halo effects.

Regarding the enlarged set of criteria used in Study 2 (Table 5a, Table 5b), the difficulty estimates reveal that, in both the RFM and the MRFM-H analyses, *structure* (part of *global impression*) was the easiest, and *description* (part of *task fulfillment*) was the most difficult criterion. Overall, under both models, the criterion difficulty and threshold estimates, respectively, were highly similar.

Table 5a: Bayesian criterion difficulty and threshold estimates, Study 2

	RFM		MRFM-H	
Criterion	Estimate	*SE*	Estimate	*SE*
Fluency				
β_1	0.06	0.09	0.21	0.12
τ_1	–2.48	0.17	–2.47	0.22
τ_2	–0.46	0.15	–0.66	0.20
τ_3	2.94	0.18	3.13	0.26
Train of thought				
β_2	0.22	0.09	0.44	0.12
τ_1	–2.48	0.16	–2.58	0.19
τ_2	0.25	0.15	0.40	0.22
τ_3	2.24	0.18	2.17	0.24
Structure				
β_3	–0.66	0.10	–0.87	0.15
τ_1	–2.61	0.18	–2.31	0.23
τ_2	–0.01	0.15	–0.04	0.19
τ_3	2.62	0.17	2.35	0.21
Completeness				
β_4	–0.53	0.09	–0.68	0.13
τ_1	–2.56	0.18	–2.55	0.32
τ_2	–0.34	0.16	–0.23	0.21
τ_3	2.90	0.17	2.79	0.27
Description				
β_5	0.62	0.09	0.92	0.14
τ_1	–2.10	0.15	–2.22	0.19
τ_2	0.08	0.16	–0.09	0.22
τ_3	2.02	0.19	2.30	0.25

Note. RFM = Rasch facets model. MRFM-H = mixture Rasch facets model for halo effects.

Table 5b: Bayesian criterion difficulty and threshold estimates, Study 2

	RFM		MRFM-H	
Criterion	Estimate	*SE*	Estimate	*SE*
Argumentation				
β_6	0.20	0.09	0.08	0.15
τ_1	–2.51	0.16	–2.42	0.22
τ_2	0.14	0.15	0.17	0.22
τ_3	2.37	0.18	2.25	0.24
Syntactic elements				
β_7	–0.29	0.08	–0.63	0.15
τ_1	–2.29	0.16	–2.54	0.22
τ_2	0.06	0.15	0.07	0.19
τ_3	2.23	0.16	2.47	0.21
Vocabulary				
β_8	–0.04	0.09	–0.07	0.13
τ_1	–2.67	0.16	–2.79	0.23
τ_2	–0.06	0.15	–0.00	0.19
τ_3	2.73	0.17	2.79	0.23
Correctness				
β_9	0.42	0.09	0.59	0.14
τ_1	–2.37	0.15	–2.47	0.21
τ_2	–0.06	0.15	–0.14	0.21
τ_3	2.44	0.18	2.61	0.25

Note. RFM = Rasch facets model. MRFM-H = mixture Rasch facets model for halo effects.

Rater parameter estimates

Looking first at the severity estimates shown in Table 6, it is evident that the models maximally agreed in each rater's location along the severity dimension; $r(18) > .99$. Thus, in both the RFM and the MRFM-H analyses, Raters 16 and 13 were the most severe raters, and Raters 1 and 7 were the most lenient raters (see also Eckes, 2015; Eckes & Jin, 2021). Regarding the rater halo estimates (last column), there was only a single rater (i.e., Rater 6) possibly subject to an illusory halo effect (*x*-estimate = .74). However, the associated standard error was very high (*SE* = .44), rendering any conclusion regarding this rater's halo tendency questionable. This finding concurs

with the negligible difference between the RFM and the MRFM-H in data–model fit (Table 2).

Table 6: Bayesian severity and halo parameter estimates for 18 raters, Study 1

	Observed scores			RFM	MRFM-H	
Rater	*N*	*M*	*SD*	α Est. (*SE*)	α Est. (*SE*)	*x* Est. (*SE*)
1	20	4.52	0.54	–2.00 (.50)	–1.99 (.50)	.00 (.04)
2	24	4.10	0.86	–1.14 (.49)	–1.14 (.50)	.00 (.00)
3	41	4.02	0.82	–1.36 (.42)	–1.34 (.43)	.00 (.00)
4	24	3.72	0.76	0.11 (.49)	0.13 (.51)	.10 (.30)
5	41	3.37	1.06	0.80 (.39)	0.80 (.41)	.00 (.00)
6	34	3.59	0.93	0.09 (.36)	0.12 (.39)	.74 (.44)
7	68	4.06	0.77	–1.82 (.37)	–1.83 (.40)	.00 (.00)
8	47	3.49	0.99	0.28 (.41)	0.32 (.40)	.40 (.49)
9	47	3.39	0.83	1.01 (.38)	1.04 (.40)	.00 (.04)
10	41	3.48	0.97	–0.62 (.38)	–0.60 (.40)	.01 (.10)
11	19	3.54	0.95	0.07 (.46)	0.11 (.48)	.02 (.14)
12	44	3.61	0.94	–0.45 (.40)	–0.42 (.40)	.00 (.00)
13	41	3.11	0.98	1.80 (.43)	1.84 (.43)	.00 (.03)
14	43	3.45	0.84	1.43 (.42)	1.47 (.42)	.01 (.08)
15	28	3.58	1.00	0.80 (.42)	0.79 (.44)	.28 (.45)
16	20	3.03	1.10	1.95 (.50)	1.97 (.49)	.06 (.23)
17	45	3.98	0.77	–0.58 (.40)	–0.55 (.41)	.00 (.00)
18	21	3.81	1.01	–0.26 (.45)	–0.26 (.48)	.00 (.00)

Note. Observed score statistics refer to the four-category rating scale ranging from 2 (*below TDN 3*) to 5 (*TDN 5*). RFM = Rasch facets model. MRFM-H = mixture Rasch facets model for halo effects. *N* is the number of essays rated. α Est. is the estimate of the rater severity parameter. *x* Est. is the estimate of the rater halo parameter.

Table 7 shows the severity and halo estimates for Study 2 raters. Again, the RFM and MRFM-H severity estimates were maximally congruent; $r(12) > .99$. For example, both models identified Rater 9 as the most severe and Rater 6 as the most lenient rater. Different from Study 1, the MRFM-H analysis suggests that there may be five raters (i.e., Raters 2, 6, 7, 10, and 11) exhibiting illusory halo tendencies. However, the associated standard errors are low enough to support such a conclusion only for three raters (i.e., Raters 6, 7, and 11).

Table 7: Bayesian severity and halo parameter estimates for 12 raters, Study 2

	Observed scores			RFM	MRFM-H	
Rater	*N*	*M*	*SD*	α Est. (*SE*)	α Est. (*SE*)	*x* Est. (*SE*)
1	30	3.39	0.77	0.60 (.27)	0.59 (.24)	.00 (.00)
2	30	3.36	1.00	0.28 (.25)	0.25 (.22)	.64 (.48)
3	30	3.41	0.94	0.12 (.25)	0.09 (.22)	.38 (.49)
4	30	3.39	0.94	–0.10 (.27)	–0.12 (.24)	.00 (.00)
5	29	3.51	0.83	0.24 (.27)	0.22 (.24)	.00 (.00)
6	30	3.83	0.88	–1.45 (.28)	–1.47 (.25)	.99 (.03)
7	30	3.53	0.93	–0.29 (.27)	–0.31 (.23)	.95 (.21)
8	30	3.36	0.95	0.07 (.26)	0.06 (.23)	.01 (.08)
9	30	3.01	0.91	0.93 (.27)	0.93 (.24)	.00 (.00)
10	30	3.49	0.87	0.07 (.26)	0.05 (.24)	.59 (.49)
11	30	3.46	1.01	0.28 (.26)	0.28 (.25)	1.0 (.00)
12	30	3.34	0.88	–0.70 (.27)	–0.73 (.25)	.02 (.15)

Note. Observed score statistics refer to the four-category rating scale ranging from 2 (*below TDN 3*) to 5 (*TDN 5*). RFM = Rasch facets model. MRFM-H = mixture Rasch facets model for halo effects. *N* is the number of essays rated. α Est. is the estimate of the rater severity parameter. *x* Est. is the estimate of the rater halo parameter.

Impact of illusory rater effects on examinee proficiency estimates

The preceding analyses revealed only weak, if at all, evidence of illusory halo effects in Study 1 and Study 2 datasets, respectively. Nonetheless, for reasons of completeness, we looked at these effects' potential impact on the examinee rank-ordering resulting from the RFM and the MRFM-H proficiency estimates.

The examinee proficiency measures estimated under the RFM and the MRFM-H were also perfectly correlated, $r(307) > .99$ (Study 1), $r(206) > .99$ (Study 2). Comparing the Study 1 RFM-based examinee rank-ordering to the examinee rank-ordering produced by the MRFM-H estimates yielded an absolute rank-order difference ranging from 0 to 7 ($M = 1.29$, $SD = 1.42$). On average, examinees' rank orderings differed by little more than one rank, depending on which model was used for estimating their proficiency. Rank differences of this magnitude do not seem to make a big difference in most practical situations (e.g., selection decisions). In Study 2, the corresponding absolute rank-order difference was even lower, ranging from 0 to 4 ($M = 0.83$, $SD = 0.87$). Thus, on average, the examinee rank orderings differed by somewhat less than one rank, depending on whether the MRFM-H or the RFM was used for estimating examinee proficiency.

Eckes and Jin (2021) applied the facets model–severity and centrality (FM-SC) to the Study 1 and Study 2 datasets in a related Bayesian modeling context. They aimed to specifically estimate centrality effects and investigate the impact of this rater effect on the examinee proficiency estimates and the resulting examinee rank-orderings. The FM-SC fit the data better than a facets model specifying a rater severity parameter only (FM-S; equivalent to the RFM in the present research). The absolute rank-order difference between estimates produced by the FM-SC and those produced by the FM-S was, on average, four ranks (Study 1) and three-and-a-half ranks (Study 2). This finding provides further evidence of the negligibly small impact of illusory halo effects, at least as far as the present datasets are concerned.

Summary and discussion

In rater-mediated assessments where each rater assigns multiple scores to examinee performances or responses, illusory rater halo effects have been notoriously difficult to pin down and separate from true halo effects (Murphy et al., 1993; Myford & Wolfe, 2004). When illusory halo effects prevail, the scores fail to differentiate accurately between the performance aspects rated. The use of correlation-based statistics (e.g., trait intercorrelations) or specifically designed rating procedures (e.g., multi-rater single-trait designs) does not seem to offer a practically viable solution to this problem (Bechger et al., 2010; Lai et al., 2016).

We opted for a Rasch measurement approach building on the facets modeling framework (Eckes, 2015; Linacre, 1989). More precisely, we applied the *mixture Rasch*

facets model for halo effects (MRFM-H; Jin & Chiu, 2022) to two real datasets that had been analyzed before in a different measurement context (Eckes & Jin, 2021). Unlike the basic Rasch facets model (RFM) that does not aim to detect halo effects, the MRFM-H distinguishes two latent classes of raters: the class of illusory halo raters and the class of normal raters showing no halo effects. For each rater, this model estimates the likelihood that he or she belongs to the class of raters exhibiting illusory halo effects.

The datasets came from two separate three-facet assessment situations with independent samples of examinees and raters using scoring rubrics, including three (Study 1) or nine criteria (Study 2). In both studies, raters assigned scores to examinees using a four-category rating scale. The three-facet rating data provided the input to analyses based on the MRFM-H. For comparison purposes, we also analyzed the Study 1 and Study 2 datasets building on the traditional RFM.

We conducted the various facets analyses adopting Bayesian MCMC estimation procedures using the R package R2jags (Su & Yajima, 2021). Three chains were run to estimate model parameters and to provide convergence diagnostics. As evidenced by the proportional scale reduction factor (PSRF) values computed for each RFM and MRFM-H parameter, respectively, the chains converged to the posterior distribution without any problem. In response to the first research question, the posterior predictive *p*-values (PPP-values) indicated that each model had satisfactorily high data–model fit.

For evaluating the model's relative fit to the data, we used three different Bayesian comparison indices (Gelman et al., 2013, 2014; Luo & Al-Harbi, 2017): the Deviance Information Criterion (DIC), the Watanabe–Akaike Information Criterion (WAIC), and the Leave-One-Out Information Criterion (LOOIC). Unlike DIC, the last two indices were computed using the R package loo (Vehtari et al., 2020). The differences in model fit were consistently lower for the RFM, but these differences were negligibly small in most cases. The only exception was that the DIC statistic favored the RFM over the MRFM-H in Study 1. Therefore, the MRFM-H did not outperform the RFM in model fit, answering the second research question. Overall, these results seem to indicate that neither the Study 1 nor the Study 2 raters were subject to illusory halo effects in any substantial way.

Responding to the third research question, we looked at each rater's likelihood to show illusory halo effects. We did so, although the model fit comparisons did not speak in favor of the MRFM-H. Also, we wanted to compensate for the model's potentially reduced power of detecting such effects when raters scored only small samples of examinee performances. In particular, the number of essays each rater scored ranged from 19 to 68 in Study 1 (M = 36.0); in Study 2, all raters scored 30 essays (except for one rater scoring 29 essays). Compared to the rater workload typical of large-scale assessments, these numbers can be considered small. Whereas in Study 1, the estimates (with their associated standard errors) of the x-parameter did not suggest any rater to exhibit illusory halo, in Study 2, we identified three raters with x-estimates between .95 and 1.0 (and small SEs), indicating a tendency to provide illusory halo

ratings. The model fit comparisons may have precisely missed these tendencies due to the low sample size. In any case, it appears worthwhile inspecting these three raters' observed rating vectors closely to learn more about their possibly aberrant rating behavior and, thus, inform rater training (or retraining) activities.

This conclusion leads us to discuss some practical implications of the present research. For assessments in the particular context of TestDaF examinations, our findings seem to suggest that illusory halo effects do not prevail in raters' judgments of examinee performances. At first sight, this is good news for TestDaF scoring rubric development and rating quality assurance measures. However, as already mentioned, in both studies, raters scored a relatively small set of examinee performances. For example, in a typical TestDaF examination, raters have to score a minimum of 63 essays (i.e., 60 essays plus three essays scored by all raters), often many more (a rater's workload may exceed 200 essays).[4] Therefore, the present finding that raters were not subject to illusory halo effects to any substantial degree needs to be validated in large-scale assessments, where raters have a much higher workload on average (Jin & Chiu, 2022). TestDaF exams administered long after the two exams we considered here were taken by many thousands of examinees and, therefore, seem well suitable for this purpose.

A related caveat refers to the magnitude of differences in criterion difficulties. In Study 1, only three criteria were used, and two of these criteria (*task fulfillment*, *linguistic realization*) had highly similar difficulty estimates. In Study 2, raters scored performances based on a rubric containing nine different criteria, but the difficulty estimates for most of these criteria were again not markedly different. Low variation in criterion difficulty, of course, decreases the likelihood of accurately classifying raters as normal raters or illusory halo raters (Jin & Chiu, 2022). The lessened classification accuracy may provide another explanation for why three halo raters were flagged in Study 2, but the Bayesian comparison statistics did not favor the MRFM-H. In future MRFM-H applications to TestDaF rating data, close attention should be paid to the extent to which criterion difficulties vary before concluding that raters are free from illusory halo effects.

For rater-mediated assessments more generally, the present findings substantiate the utility of Jin and Chiu's (2022) mixture Rasch facets model. Standing in the strong tradition of facets modeling, the MRFM-H provides estimates of examinee proficiency corrected for rater severity differences and corrected for illusory halo effects (if any). Bayesian parameter estimation, conducted with free JAGS software within the R environment, proved to be efficient at fitting the data to the model. Therefore, the MRFM-H holds much promise to contribute significantly to achieving high rating quality.

[4] These figures refer to the paper-based version of the TestDaF only. As of autumn 2021, the completely web-based (digital) TestDaF was released (Kecker et al., 2022). The digital TestDaF, which will eventually replace the paper-based version, employs a holistic scoring rubric where halo effects are minimized by design.

Finally, in this research, we focused exclusively on applying a new psychometric model to address the issue of detecting halo. We did not focus on the cognitive or judgmental processes that may cause illusory halo effects, if observed, in the first place. Any of the processes discussed by Fisicaro and Lance (1990) may be the objective of future research. Such research would preferably combine Rasch-based measurement approaches such as the one presented here with more qualitatively oriented approaches building on structured interviews, stimulated recall, or verbal protocol analysis (Myford, 2012; Turner, 2014).

Conclusion

Being one of the four classic rater effects (severity/leniency, halo, central tendency, and restriction of range), “the halo effect has been the most studied and has received the widest attention in the research literature” (Myford & Wolfe, 2003, p. 395). The wide research interest notwithstanding, psychometric methods and statistics aiming at detecting or measuring these sources of error in human judgment have met with little success. By contrast, methods for measuring other classic rater effects, particularly severity/leniency and central tendency (Eckes, 2015, 2019; Eckes & Jin, 2021; Engelhard & Wind, 2018), have been much more successful. The long-standing problem with halo effect detection is distinguishing true from illusory halo. More than 100 years after Thorndike (1920) coined the term "halo", the MRFM-H (Jin & Chiu, 2022) is an important step forward in separating these two components and, therefore, measuring, and controlling for, illusory halo effects. Future MRFM-H-based research using various kinds of rating designs, scoring rubrics, and rater and examinee samples will help broaden our understanding of the nature of halo effects and their role in rater-mediated assessments.

References

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

Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology*, *77*(6), 975–985. https://doi.org/10.1037/0021-9010.77.6.975

Bartlett, C. J. (1983). What’s the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. *Journal of Applied Psychology*, *68*(2), 218–226. https://doi.org/10.1037/0021-9010.68.2.218

Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement*, *34*(8), 607–619. https://doi.org/10.1177/0146621610367897

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*(2), 218–244. https://doi.org/10.1037/0033-2909.90.2.218

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. https://doi.org10.3726/978-3-653-04844-5

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*, *21*(3-4), 131–153. 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. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Erlbaum.

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

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, *14*(4), 419–429. https://doi.org/10.1177/014662169001400407

Fisicaro, S. A., & Vance, R. J. (1994). Comments on the measurement of halo. *Educational and Psychological Measurement*, *54*(2), 102–125. https://doi.org/10.1177/0013164494054002010

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153–160. https://doi.org/10.1080/01621459.1979.10481632

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.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

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

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

Jin, K.-Y., & Chiu, M. M. (2022). A mixture Rasch facets model for rater's illusory halo effects. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-021-01721-3

Jin, K.-Y., & Eckes, T. (2021). Detecting differential rater functioning in severity and centrality: The dual DRF facets model. *Educational and Psychological Measurement.* Advance online publication. https://doi.org/10.1177/00131644211043207

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.*

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

Kecker, G., Zimmermann, S., & Eckes, T. (2022). Der Weg zum digitalen TestDaF: Konzeption, Entwicklung und Validierung [The road to the digital TestDaF: Conceptualization, development, and validation]. In P. Gretsch & N. Wulff (Eds.), *Deutsch als Zweit- und Fremdsprache in Schule und Beruf* (pp. 393–410). Schöningh.

Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox.

Lai, E. R., Wolfe, E. W., & Vickers, D. (2015). Differentiation of illusory and true halo in writing scores. *Educational and Psychological Measurement*, *75*(1), 102–125. https://doi.org/10.1177/0013164414530990

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

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

Linacre, J. M. (2000). Comparing and choosing between "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). *Rasch Measurement Transactions*, *14*(3), 768. https://www.rasch.org/rmt/rmt143k.htm

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

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

Luo, Y., & Al-Harbi, K. (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychological Test and Assessment Modeling*, *59*(2), 183–205. https://www.psychologie-aktuell.com/journale/psychological-test-and-assessment-modeling/currently-available/inhaltlesen/2017-2.html

Marais, I., & Andrich, D. (2011). Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, *12*(3), 194–211. http://jampress.org/pubs.htm

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., & Adams, R. J. (2000). The implications of halo effects and item dependencies for objective measurement. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 243–257). Ablex.

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

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, *84*(3), 802–829. https://doi.org/10.1007/s11336-019-09679-0

Murphy, K. R. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology*, *67*(2), 161–164. https://doi.org/10.1037/0021-9010.67

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, *78*(2), 218–225. https://doi.org/10.1037/0021-9010.78.2.218

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice*, *31*(3), 48–49. https://doi.org/10.1111/j.1745-3992.2012.00243.x

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

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

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

Plummer, M. (2017). *JAGS version 4.3.0 user manual*. https://people.stat.sc.edu/hansont/stat740/jags_user_manual.pdf

Pulakos, E. D., Schmitt, N., & Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology*, *71*(1), 29–32. https://doi.org/10.1037/0021-9010.71.1.29

R Core Team (2021). *R: A language and environment for computing* (Version 4.1.2) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

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

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), 583–616. https://doi.org /10.1111/1467-9868.00353

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B*, *76*(3), 485–493. https://doi.org/10.1111/rssb.12062

Su, Y.-S., & Yajima, M. (2021). *Package 'R2jags'* (Version 0.7-1) [Computer software]. https://cran.r-project.org/web/packages/runjags/index.html

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Turner, C. E. (2014). Mixed methods research. In A. J. Kunnan (Ed.), *The companion to language assessment: Evaluation, methodology, and interdisciplinary themes* (Vol. 3, pp. 1403–1417). Wiley.

van der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica*, *59*(1), 45–56. https://doi.org/10.1111/j.1467-9574.2005.00278.x

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., Gelman, A., Goodrich, B., Piironen, J., & Nicenboim, B. (2020). *Package 'loo'* (Version 2.4.1) [Computer software]. https://cran.r-project.org/web/packages/loo/index.html

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(116), 3571–3594. https://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf

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

Wolfe, E. W., & 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.