



# Examining Illusory Halo Effects Across Successive Writing Assessments: An Issue of Stability and Change


Thomas Eckes   
*TestDaF Institute, University of Bochum*


Kuan-Yu Jin   
*Hong Kong Examinations and Assessment Authority*

Halo effects are a common source of rating errors in assessments. When raters assign scores to examinees on multiple performance dimensions or criteria, they may fail to discriminate between them, lowering each criterion's information value regarding an examinee's performance. Using the mixture Rasch facets model for halo effects (MRFM-H), we studied halo tendencies in four successive high-stakes writing assessments administered to 15,677 examinees spanning 10 months involving 162 raters. The MRFM-H allows separating between illusory halo due to judgmental biases and true halo due to the actual overlap between the criteria. Applying this model, we aimed to detect illusory halo effects in the first exam, tracking the effects' occurrence across the other three exams. We also ran the standard Rasch facets model (RFM) and computed raw-score correlational and standard deviation halo indices,  $r_H$  and  $SD_H$ , for comparison purposes. The findings revealed that (a) the MRFM-H fit the rating data better than the RFM in all four exams, (b) in the first exam, 11 out of 100 raters exhibited illusory halo effects, (c) the halo raters showed evidence of both stability and change in their rating tendencies over exams, (d) the non-halo raters mostly remained stable, (e) the  $r_H$  and  $SD_H$  statistics did not separate between the halo and non-halo raters, and (f) the illusory halo effects had a small but demonstrable impact on examinee rank orderings, which may have consequences for selection decisions. The discussion focuses on the model's practical implications for performance assessments, such as rater training, monitoring, and selection, and highlights future research perspectives.

*Keywords:* rater effects, illusory halo, writing assessment, Rasch measurement, mixture Rasch model

---

Dr. Thomas Eckes  <https://orcid.org/0000-0002-8820-5902>

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

Requests for reprints should be sent to Dr. Thomas Eckes, TestDaF Institute, University of Bochum, Universitätsstr. 134, 44799 Bochum, Germany; e-mail: [eckes@gast.de](mailto:eckes@gast.de) or

Dr. Kuan-Yu Jin, Hong Kong Examinations and Assessment Authority, 3/F, 68 Gillies Avenue South, Kowloon City, Kowloon, Hong Kong; e-mail: [kyjin@hkeaa.edu.hk](mailto:kyjin@hkeaa.edu.hk)

Halo effects refer to a common source of error in rater-mediated assessments (Cooper, 1981; Saal et al., 1980). When raters assign multiple scores to examinees, halo effects manifest in similar scores on conceptually distinct criteria (Myford & Wolfe, 2004). For example, raters using an analytic scoring rubric to evaluate examinees' writing performances may give unduly similar scores for content, organization, and language.

Halo effects typically arise from biased cognitive or judgmental processes (Fisicaro & Lance, 1990). Thus, a first impression of an examinee based on performance-irrelevant cues (e.g., nice handwriting) may dominate a rater's perception of the examinee's performance. Alternatively, a single, subjectively salient performance feature (e.g., grammatical correctness) may guide the ratings on the whole set of criteria. Finally, raters may fail to discriminate adequately between the criteria. No matter what process is involved in any given instance, halo effects reduce the amount of information on the performance each criterion score conveys. The result is an erroneous evaluation of an examinee's performance, threatening the validity and fairness of the assessment outcomes (Knoch et al., 2021; Wind & Peterson, 2018; Wolfe & Song, 2016).

Over the decades since Thorndike (1920) first coined the term "halo," many psychometric methods and statistics for halo detection have been proposed. More commonly used statistics rest on the observed ratings or raw scores, such as criterion intercorrelations, the  $r_H$  statistic hereafter, or within-examinee standard deviations, the  $SD_H$  statistic (Fisicaro & Vance, 1994; Murphy, 1982; Pulakos et al., 1986; Saal et al., 1980). Following these statistics' rationale, raters subject to halo effects show heightened correlations between criteria or reduced within-examinee variability across criteria. Other researchers discussed halo detection methods based on psychometric models, particularly many-facet Rasch models (Engelhard, 1994; Kim, 2020; McNamara & Adams, 2000; Myford & Wolfe, 2004; O'Grady,

2023). Under such a measurement approach, halo effects are typically assumed to be associated with mean-square rater fit statistics less than 1 (overfit), indicating less score variability than expected based on the model. However, rater fit statistics often do not provide conclusive evidence of halo effects. Myford and Wolfe (2004) pointed out that mean-square fit statistics, in the presence of halo effects, may be smaller or greater than 1, depending on the assessment context, such as the extent to which criterion or trait difficulties vary. Therefore, such statistics are generally not recommended for routine use (e.g., Wolfe & Song, 2016); they will not be discussed here further.

More importantly, previous raw-score and measurement approaches fail to distinguish between true and illusory halo (Bartlett, 1983; Murphy et al., 1993). Many criteria used in analytic scoring correlate with each other to some extent because they relate to the same construct or performance the assessment targets. This kind of intrinsic relation among traits or criteria gives rise to a true (or valid) halo. True halo contrasts with illusory or (invalid) halo, which constitutes the real issue because it reflects raters' cognitive biases and distortions.

To identify illusory halo effects and separate these from true halo, researchers have proposed using special rating designs, for example, designs involving a group of expert raters as a reference or standard (Balzer & Sulsky, 1992). Others suggested comparing single raters, each scoring multiple criteria, with multiple raters scoring a single criterion (Bechger et al., 2010; Lai et al., 2015). As a significant drawback, such designs are typically challenging to implement, cost-intensive, and possibly less effective than intended; for example, expert ratings are not readily available in many applied settings and may be subject to an illusory halo for their part.

The present study adopted a specific measurement approach to separate true from illusory halo effects—the mixture Rasch facets model for halo effects (MRFM-H; Jin & Chiu, 2022). Besides estimating examinee proficiency,

rater severity, and criterion difficulty, as in the standard Rasch facets model (RFM; Linacre, 1989), the MRFM-H allows the detection of raters subject to illusory halo.

We used the MRFM-H to study halo effects in four successive writing assessments administered across 10 months in a high-stakes context. This longitudinal perspective allowed us to examine how much halo effects remained stable or changed over time. Following research highlighting the variability of rater characteristics such as severity, centrality, or consistency over time (e.g., Congdon & McQueen, 2000; Hoskens & Wilson, 2001; Lamprianou et al., 2021; Lim, 2011; Lunz et al., 1996; Myford & Wolfe, 2009; Uto, 2023), we expected to find some evidence for variability of illusory halo effects among operational raters.

### Measuring Illusory Halo Effects

The MRFM-H (Jin & Chiu, 2022) extends the standard RFM, incorporating two rater-related entities: (a) a latent rater severity dimension and (b) two latent classes of raters, where the first class consists of "normal" (or "non-halo") raters and the second class consists of "(illusory) halo raters." Specifically, the MRFM-H includes an illusory halo parameter, defined as a binary variable following a Bernoulli distribution, indicating the latent class membership of a particular rater. Considering this definition, the model stands in the tradition of mixture Rasch measurement (Jin & Wang, 2017; Rost, 1990, 2001; von Davier & Rost, 2016), where a person's class membership is unknown and must be estimated from the data. The MRFM-H is given as follows:

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

In Equation 1,  $P_{nikj}$  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  $j$ ,  $\theta_n$  is the proficiency of

examinee  $n$ ,  $\beta_i$  is the difficulty of criterion  $i$ ,  $\alpha_j$  is the severity of rater  $j$ , and  $\tau_{ik}$  is the difficulty of receiving on criterion  $i$  a rating of  $k$  relative to  $k - 1$  (rating scale category threshold or "step difficulty").

The examinee, criterion, rater, and threshold parameters (Equation 1) are identical to the usual RFM specification. Unlike the RFM, however, the MRFM-H also contains the three terms  $x_j$ ,  $\beta^*$ , and  $\tau_k^*$ . The meaning of these terms is explained below.

The  $x_j$  term represents the (illusory) halo parameter. More precisely, this parameter is a binary variable following a Bernoulli distribution  $\pi$ , where  $x_j$  provides the latent class membership of a particular rater. Thus, rater  $j$  belongs (a) to the class of (illusory) halo raters, indicated by  $x_j = 1$ , or (b) to the class of non-halo raters, indicated by  $x_j = 0$ . Hence, the likelihood of the observed ratings under the MRFM-H reveals the latent class to which rater  $j$  belongs. When  $x_j = 0$  for each rater  $j \in J$ , all raters are non-halo raters, and the MRFM-H reduces to the RFM (Jin & Chiu, 2022).

Finally, the  $\beta^*$  and the  $\tau_k^*$  terms designate the difficulty and threshold values of a *generalized criterion*:  $\beta^* = \beta_1 = \dots = \beta_I$ , and  $\tau_k^* = \tau_{1k} = \dots = \tau_{Ik}$ . The generalized criterion represents a combination of the individual (analytic) criteria when raters do not distinguish between them. As a result, raters give each examinee highly similar scores across the different criteria (Myford & Wolfe, 2004).

Like the RFM, the MRFM-H quantifies raters' severity tendencies and measures their impact on the assessment outcomes. However, the MRFM-H goes beyond the RFM in that it also detects the presence of illusory halo effects. It allows researchers to control their influence on the ratings statistically, simultaneously correcting the scores raters assign to examinees for severity and halo effects. That is, the examinee proficiency estimates are adjusted for the differences in the level of rater severity and the potential impact of halo effects. Thus, the MRFM-H further helps improve the

assessment's validity and fairness.

Jin and Chiu (2022) demonstrated the efficiency of the MRFM-H for detecting illusory halo in a simulation study. Classifying raters as halo raters when the estimate of the  $x_j$  parameter exceeded .50<sup>1</sup>, they showed that applying the MRFM-H yielded over 99% recovery of raters' normal versus illusory halo class membership. When this model was fitted to data containing no illusory halo effects (i.e., when the RFM was the true model), the MRFM-H recovered parameter estimation (rater severity, criterion difficulty, and thresholds) just as well. Furthermore, using the .50 cut-off, Jin and Chiu analyzed three real datasets, including one from an English essay writing assessment. Across these analyses, the MRFM-H proved valuable for identifying raters subject to illusory halo effects.

Eckes and Jin (2022) used the MRFM-H in two studies, each comprising a real dataset from small-scale writing assessments. The first dataset had 18 raters and 307 examinees, while the second dataset had 12 raters and 206 examinees. On average, raters scored 36 essays in Study 1 and 30 in Study 2. Data-model fit statistics indicated that the MRFM-H fit the data well but not better than the RFM. In line with this finding, neither study identified raters exhibiting illusory halo effects.

Based on these findings, Eckes and Jin (2022) discussed two factors potentially having an impact on the accuracy of classifying raters as normal raters or raters subject to illusory halo: (a) the number of performances each rater scored and (b) the magnitude of the differences between criterion difficulties. They assumed the classification accuracy to be lower when the samples of scored performances are small and the criteria have similar difficulty measures. Drawing on a much larger dataset, with raters scoring more performances on average, the present study aimed to probe into the occurrence

of illusory halo tendencies.

### Research Questions

Across four large-scale writing assessments successively administered over 10 months, we investigated the extent to which operational raters were subject to illusory halo effects. For this purpose, we built on Jin and Chiu's (2022) mixture Rasch facets model for halo effects (MRFM-H). Using the same data, we ran the standard Rasch facets model (RFM) and computed raw-score  $r$  and  $SD$  halo statistics to compare with the MRFM-H findings. Based on the results from these analyses, the present research addressed the following questions:

1. How does the MRFM-H compare to the RFM regarding data-model fit across the four writing assessments?
2. Are individual raters exhibiting illusory halo effects in the first assessment?
3. Does the raters' latent class membership (i.e., halo or non-halo) identified in the first assessment change or remain stable across the other three assessments?
4. Do the raw-score  $r$  and  $SD$  halo statistics differentiate between the halo and non-halo raters?
5. How do the illusory halo effects impact the examinee rank order derived from the writing proficiency estimates?

### Method

#### Participants

All 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. On average, raters scored between 48 and 55 essays. Table 1 presents detailed information on the examinee and rater samples participating in the four exams administered from November 2019 (Exam 1) to September 2020 (Exam 4).

#### Instruments and Procedure

**Table 1**

*Characteristics of Four Writing Examinations*

Characteristic	Exam 1	Exam 2	Exam 3	Exam 4
Administration	Nov. 2019	Febr. 2020	June 2020	Sept. 2020
Facets				
Examinees ( $N$ )	5,191	5,501	2,619	2,366
Raters ( $J$ )	100 <sup>a</sup>	107	55	52
Criteria ( $I$ )	9	9	9	9
Essays rated				
Min	16	23	25	25
Max	145	100	70	65
$M$	54.9	54.4	50.6	48.4
$SD$	29.0	24.6	13.0	13.2

*Note.* The ratings were assigned in each exam according to a sparse assessment network (performance links design). Several raters participated in more than one exam. In Exams 3 and 4, the numbers of examinees and raters were much lower due to the COVID-19 pandemic.

<sup>a</sup> In the Exam 1 rater group (the reference sample), 11 raters exhibited illusory halo effects, 89 raters were placed in the non-halo class; of these non-halo raters, 13 raters also completed essay ratings in the other three exams (the comparison sample).

The writing task was part of the Test of German as a Foreign Language (TestDaF, *Test Deutsch als Fremdsprache*) in its paper-based version—an officially recognized language exam for international students (Eckes & Althaus, 2020; Norris & Drackert, 2018)<sup>2</sup>. Examinee performances in four test sections (reading, listening, writing, and speaking) are related to one of three increasingly higher levels of language proficiency, the TestDaF levels (*TestDaF-Niveaus*, TDNs). The writing section uses a single task and assesses an examinee's ability to produce a coherent and well-structured text on a given topic taken from the academic context.

In each writing assessment, the rating design corresponded to the performance links design Jin and Chiu (2022) used in their simulation study; that is, the design was incomplete but connected (DeMars et al., 2023;

Eckes, 2023b; Wind & Ge, 2021). A single rater scored each examinee's performance (essay), and all raters scored a common set of three "linking" performances to establish connectivity between raters. For example, a rater scoring a unique set of 57 essays in Exam 1 also had to score each linking essay. The links were selected from a larger set of essays obtained in an extensive pre-test of the writing task. Together, the linking performances represented the range of typical writing proficiency levels from low to high. Each time, a different pre-tested writing task was used to select linking performances for use with Exams 1 to 4.

Across the exams, raters used the same analytic scoring rubric, comprising a four-category scale, the TDN scale, with categories *below TDN 3*, *TDN 3*, *TDN 4*, and *TDN 5*. For computation purposes, *below TDN 3* was scored "2," and the other levels were scored from "3" to "5." Essay ratings were provided separately on nine criteria: *fluency*, *train of thought*, and *structure* (representing the higher-level criterion *global impression*); *completeness*, *description*, and *argumentation* (representing *task*

<sup>1</sup> The .50 cut-off value is common in two-class mixture modeling, indicating that a person is more likely to belong to one class than another (e.g., Baghaei & Carstensen, 2013; Kreitchmann et al., 2023).

<sup>2</sup> A completely revised, web-based version, the digital TestDaF, was released in late 2020 and will eventually replace the paper-based TestDaF version (g.a.s.t., 2020). The present research dealt exclusively with the paper-based version.

fulfillment); and *breadth of syntactic elements*, *vocabulary*, and *correctness* (representing linguistic realization).

Data Analysis

Across exams, we adopted the same methodological approach. Specifically, we used Bayesian data analysis procedures (Gelman et al., 2013; Kruschke, 2015; Lunn et al., 2013) with Markov chain Monte Carlo (MCMC) parameter estimation implemented in the JAGS freeware (JAGS = Just Another Gibbs Sampler; Plummer, 2017). In addition, 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, 2022).

We specified the identical prior distributions of the model parameters as in Jin and Chiu (2022; see also Eckes & Jin, 2022):

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

$\beta_i \sim N(0, 0.1),$  (3)

$\alpha_j \sim N(0, 0.1),$  (4)

$\tau_{ik} \sim N(0, 0.1),$  (5)

$x_j \sim \text{Bernoulli}(\pi),$  (6)

where  $N(\mu, \tau)$  is the normal distribution with mean  $\mu$  and precision  $\tau$ , for  $\tau > 0$ ; the variance  $\sigma^2$  of the normal distribution is  $1/\tau$ ;  $\text{Bernoulli}(\pi)$  is the Bernoulli distribution with probability  $\pi$  (Lunn et al., 2013). Also, we used the following priors for the hyperparameters:

$\mu \sim N(0, 0.1),$  (7)

$\sigma^2 \sim \text{Gamma}(0.1, 0.1),$  (8)

$\pi \sim \text{Beta}(1, 1),$  (9)

where  $\text{Gamma}(r, \lambda)$  is the Gamma distribution with shape  $r$  and rate  $\lambda$ ;  $\text{Beta}(\alpha, \beta)$  is the Beta distribution with shape parameters  $\alpha$  and  $\beta$  (Lunn et al., 2013).

Three MCMC chains were run to assess convergence to the posterior distributions. 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 of a given parameter; similarly, the posterior standard deviation was used to estimate 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).

We used the Gelman-Rubin statistic’s potential scale reduction factor (PSRF; Gelman & Rubin, 1992) as an index of convergence to the posterior distribution. It is commonly suggested to infer that the chains have converged if the PSRF values are close to 1 (i.e.,  $\text{PSRF} < 1.1$ ; 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). Extreme posterior predictive  $p$ -values (PPP values), that is, values close to 0 or 1, indicate insufficient 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, we employed three different criteria. The first criterion was the Bayesian deviance information criterion (DIC; Spiegelhalter et al., 2002, 2014; van der Linde, 2005). Models showing smaller DIC values generally fit better (Levy & Mislevy, 2016, p. 248). We also computed two other Bayesian model comparison statistics: 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, the WAIC and LOOIC statistics require using the whole posterior distribution instead of point estimates, which is why these two criteria are viewed as fully Bayesian (AlHakmani & Sheng, 2023; Gelman et al., 2013, 2014; Luo & Al-Harbi, 2017).

Table 2  
Bayesian RFM and MRFM-H Fit and Comparison Statistics in Four Exams

Statistic	Exam 1		Exam 2		Exam 3		Exam 4	
	RFM	MRFM-H	RFM	MRFM-H	RFM	MRFM-H	RFM	MRFM-H
PSRF (min–max)								
Examinee proficiency	1.000–1.035	1.000–1.064	1.000–1.036	1.000–1.040	1.000–1.108	1.000–1.067	1.000–1.038	1.000–1.024
Rater severity	1.000–1.046	1.000–1.016	1.001–1.035	1.000–1.060	1.000–1.017	1.001–1.030	1.000–1.066	1.000–1.033
Criterion difficulty	1.000–1.004	1.001–1.004	1.001–1.005	1.000–1.006	1.000–1.005	1.000–1.005	1.000–1.003	1.000–1.004
Thresholds	1.000–1.006	1.000–1.004	1.000–1.007	1.000–1.003	1.000–1.007	1.000–1.006	1.000–1.005	1.000–1.002
PPP-value	.517	.500	.326	.328	.449	.445	.576	.580
DIC	87,074.5	86,539.3	86,655.3	86,533.2	41,127.5	41,004.4	37,084.6	37,056.5
WAIC	85,800.0	85,575.5	85,958.8	85,560.4	40,715.5	40,646.0	36,671.9	36,613.0
LOOIC	85,916.8	85,691.5	86,089.6	85,692.1	40,774.9	40,709.2	36,725.7	36,670.1

Note. Throughout the RFM and MRFM-H analyses, the partial credit versions were used. PSRF = Potential scale reduction factor. PPP-value = Posterior predictive  $p$ -value. DIC = Deviance information criterion. WAIC = Watanabe–Akaike information criterion. LOOIC = Leave-one-out information criterion. Thresholds are Rasch–Andrich thresholds.

Results

Data–Model Fit

Table 2 summarizes the Bayesian convergence and data–model fit statistics. 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.

Across the four exams, the three model comparison statistics (DIC, WAIC, and LOOIC) unanimously favored the MRFM-H over the RFM. This finding attests to non-negligible illusory halo effects in each exam.

Rater Halo Statistics and Parameter Estimates

As explained above, raters with halo parameter estimates ( $x$  estimates) greater than .50 were classified as halo raters; all the others were classified as non-halo raters (Jin & Chiu, 2022, p. 2754). Following this rule, of the 100 Exam 1 raters, 11 raters proved to be subject to illusory halo effects, and 89 raters were placed

in the non-halo class. Among the non-halo raters in Exam 1, five raters had  $x$  estimates ranging from .01 to .14; the remaining 84 raters had  $x$  estimates equal to .00. The relative frequencies of halo raters in the remaining three exams were 8.4% (Exam 2), 10.9% (Exam 3), and 11.5% (Exam 4). Thus, some 10% of the raters were classified as halo raters across exams.

Using the sample of Exam 1 raters as a reference, we identified 13 non-halo raters who also completed essay ratings in the other three exams. These raters formed the comparison sample across the exams. Notably, the Exam 1 halo and non-halo rater samples allowed us to track the presence or absence of illusory halo effects over the exams considered here.

Table 3 presents detailed results for the Exam 1 raters. This table gives raw-score summary statistics, raw-score halo statistics, and rater parameter estimates under the RFM and the MRFM-H for the 11 halo raters, H01 to H11 (upper panel), and the 13 non-halo raters, N01 to N13 (lower panel).

As shown in the last column, most halo raters had  $x$  estimates equal to or close to 1.0. However, only one rater, Rater H03, had an  $x$  estimate slightly above the cut-off value of .50. Compared to the other estimates, the associated standard error was high ( $SE = .50$ ), indicating a

**Table 3**  
*Observed Rater Statistics and Bayesian RFM and MRFM-H Parameter Estimates for Halo and Non-Halo Raters in Exam 1*

Rater	Observed scores			Halo statistics		RFM	MRFM-H	
	$N_R$	$M$	$SD$	$r_H$	$SD_H$	$\alpha$ Est. (SE)	$\alpha$ Est. (SE)	$x$ Est. (SE)
Halo raters <sup>a</sup>								
H01	51	3.56	.70	.56	.59	−0.13 (.23)	−0.15 (.23)	1.0 (.06)
H02	70	3.32	.53	.49	.56	−0.05 (.22)	−0.08 (.20)	1.0 (.0)
H03	24	3.10	.44	.43	.55	0.74 (.27)	0.72 (.26)	.54(.50)
H04	30	3.65	.64	.57	.53	−0.64 (.26)	−0.63 (.26)	.95 (.21)
H05	90	3.41	.58	.59	.54	−0.16 (.19)	−0.18 (.20)	1.0 (.0)
H06	60	3.67	.46	.41	.55	−0.62 (.22)	−0.64 (.21)	1.0 (.0)
H07	80	2.81	.63	.60	.46	1.64 (.20)	1.58 (.20)	1.0 (.05)
H08	60	3.50	.85	.74	.48	−0.21 (.22)	−0.22 (.22)	1.0 (.0)
H09	29	3.29	.54	.54	.54	0.27 (.26)	0.23 (.26)	.91 (.29)
H10	25	3.32	.68	.62	.55	0.06 (.27)	0.06 (.26)	.86 (.34)
H11	30	3.36	.56	.54	.50	0.08 (.25)	0.06 (.25)	1.0 (.03)
Non-halo raters <sup>b</sup>								
N01	60	3.43	.50	.50	.57	−0.65 (.21)	−0.67 (.21)	.0 (.0)
N02	139	2.96	.44	.38	.55	1.22 (.17)	1.20 (.16)	.0 (.0)
N03	90	3.35	.53	.60	.47	−0.19 (.19)	−0.22 (.20)	.0 (.0)
N04	60	3.54	.67	.73	.39	−0.20 (.21)	−0.22 (.21)	.0 (.0)
N05	59	3.58	.66	.59	.57	−0.94 (.22)	−0.97 (.22)	.0 (.0)
N06	30	3.88	.63	.62	.56	−0.86 (.25)	−0.87 (.26)	.0 (.0)
N07	60	3.66	.60	.61	.49	−0.95 (.22)	−0.97 (.22)	.0 (.0)
N08	100	3.67	.93	.76	.51	−0.93 (.19)	−0.97 (.18)	.0 (.0)
N09	100	3.04	.67	.65	.53	1.28 (.19)	1.28 (.19)	.0 (.0)
N10	25	3.56	.45	.39	.61	−0.13 (.26)	−0.15 (.26)	.01 (.08)
N11	25	3.70	.55	.42	.62	−0.33 (.26)	−0.36 (.27)	.0 (.0)
N12	30	2.95	.62	.61	.58	1.53 (.26)	1.55 (.26)	.0 (.0)
N13	50	3.02	.80	.68	.53	1.08 (.23)	1.07 (.23)	.00 (.03)

Note. Observed score statistics refer to the four-category rating scale ranging from 2 (*below TDN* 3) to 5 (*TDN* 5).  $N_R$  = number of essays rated.  $r_H$  = halo correlation statistic.  $SD_H$  = halo  $SD$  statistic. RFM = Rasch facets model. MRFM-H = mixture Rasch facets model for halo effects.  $\alpha$  Est. = rater severity parameter estimate.  $x$  Est. = rater halo parameter estimate.

<sup>a</sup> “Halo raters” are raters with  $x$  estimates  $> .50$  in Exam 1. <sup>b</sup> “Non-halo raters” are raters with  $x$  estimates  $\leq .50$  in Exam 1 and scoring essays in the other three exams.

considerable amount of uncertainty around this rater’s class membership. Therefore, Rater H03 can be considered a borderline case.

Much as expected, the raw-score rater halo statistics,  $r_H$  and  $SD_H$ , were negatively correlated across classes,  $r(24) = -.61, p < .01$ , and within classes: halo raters,  $r(11) = -.51, ns$ , non-halo raters,  $r(13) = -.66, p < .05$ . Neither statistic was significantly correlated with the  $x$  estimates across raters from both classes. For  $r_H$ ,  $r(24) = -.08, ns$ ; for  $SD_H$ ,  $r(24) = -.07, ns$ .

Between-class comparisons confirmed that the  $r_H$  and  $SD_H$  statistics did not differentiate between the halo and non-halo raters. For the  $r_H$  statistic,  $t(22) = -0.59, ns$ ; for the  $SD_H$  statistic,  $t(22) = -0.24, ns$ .

Finally, looking at the severity estimates shown in Table 3, it is evident that the models maximally agreed in each rater’s location along the severity dimension; the correlation between the severity estimates was close to 1,  $r(24) > .99$ . This finding concurs with the negligible

difference between the RFM and the MRFM-H in data–model fit statistics, PSRF and PPP values (Table 2).

**Criterion Difficulty and Threshold Parameter Estimates**

Figure 1 displays the RFM and the MRFM-H estimates of criterion difficulty and criterion-specific thresholds in Exam 1. The two models generally yielded almost identical criterion estimates. Thus, *structure* and *syntactic elements* (Criterion 3 and 7) were the easiest in the RFM and the MRFM-H analyses, and *argumentation* and *correctness* (Criterion 6 and 9) were the most difficult criteria. Finally, the close correspondence of threshold estimates

across criteria indicated that the Exam 1 raters used and interpreted the rating scale in much the same way, irrespective of the criterion considered. The results for the other three exams (not shown here) looked the same: close correspondence between RFM and MRFM-H estimates, similar criterion locations along the difficulty scale, and highly stable rating scale structure across criteria.

**Stability and Change in Raters’ Halo Tendencies Across Exams**

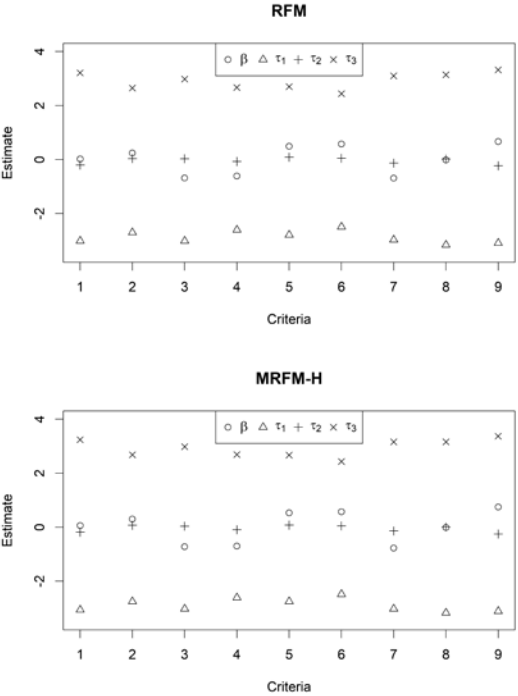
Table 4 presents the halo parameter estimates for the class of 11 halo raters and the comparison sample of 13 non-halo raters across the four exams. Looking first at the estimates for the 10 halo raters who participated in at least two exams (Rater H04 participated in Exam 1 only), there was evidence for stability and change. Three raters, H01, H02, and H09, constantly exhibited illusory halo effects in all two or three exams in which they participated. Another three raters, H05, H08, and H11, first constantly exhibited halo in two or three exams but then joined the group of non-halo raters in the last exam. Similarly, three other raters, H06, H07, and H10, changed from halo to non-halo group membership once. Finally, the borderline case, H03, exhibited a weak halo tendency in the first exam but no halo effects in the other two exams in which he or she participated.

Regarding the group of non-halo raters, there was unequivocal evidence for the stability of non-halo tendencies across exams, with one exception. Rater N13 started in Exam 1, showing no halo effect. However, this rater exhibited halo effects in Exam 2, returned to non-halo tendencies in Exam 3, and ended up with halo tendencies again in Exam 4.

**Impact of Illusory Halo Effects on Examinee Proficiency Estimates**

The examinee proficiency measures estimated under the RFM and the MRFM-H were almost perfectly correlated in each exam. However, viewed from the perspective of norm-referenced assessments that report examinees’

**Figure 1**  
*Criterion Difficulty ( $\beta$ ) and Threshold Parameter ( $\tau_1, \tau_2, \tau_3$ ) Estimates Under the RFM (Upper Panel) and the MRFM-H (Lower Panel) in Exam 1*



Note. The criteria were (1) fluency, (2) train of thought, (3) structure, (4) completeness, (5) description, (6) argumentation, (7) syntactic elements, (8) vocabulary, and (9) correctness.

**Table 4**  
*MRFM-H Halo Parameter Estimates for Halo and Non-Halo Raters in Four Exams*

Rater	Exam 1			Exam 2			Exam 3			Exam 4		
	$N_R$	$x$ Est.	SE	$N_R$	$x$ Est.	SE	$N_R$	$x$ Est.	SE	$N_R$	$x$ Est.	SE
Halo raters <sup>a</sup>												
H01	51	1.0	.06	60	1.0	.0	60	.97	.16	–	–	–
H02	70	1.0	.0	70	1.0	.0	–	–	–	60	.99	.08
H03	24	.54	.50	25	.0	.03	–	–	–	28	.0	.0
H04	30	.95	.21	–	–	–	–	–	–	–	–	–
H05	90	1.0	.0	100	1.0	.0	60	.99	.12	60	.0	.0
H06	60	1.0	.0	100	.0	.0	60	.0	.0	–	–	–
H07	80	1.0	.05	70	.07	.25	–	–	–	–	–	–
H08	60	1.0	.0	60	1.0	.0	60	1.0	.03	60	.26	.44
H09	29	.91	.29	25	.99	.08	–	–	–	–	–	–
H10	25	.86	.34	–	–	–	40	.0	.0	40	.01	.12
H11	30	1.0	.03	60	1.0	.0	60	.0	.0	60	.0	.0
Non-halo raters <sup>b</sup>												
N01	60	.0	.0	100	.0	.0	60	.0	.0	42	.0	.0
N02	139	.0	.0	90	.0	.0	60	.05	.23	60	.0	.0
N03	90	.0	.0	100	.0	.0	57	.0	.0	60	.0	.0
N04	60	.0	.0	60	.0	.0	60	.0	.0	60	.0	.0
N05	59	.0	.0	60	.0	.0	50	.0	.0	40	.0	.0
N06	30	.0	.0	90	.0	.0	60	.0	.0	60	.0	.0
N07	60	.0	.0	60	.0	.0	60	.0	.0	60	.0	.0
N08	100	.0	.0	80	.0	.0	60	.0	.0	60	.0	.0
N09	100	.0	.0	70	.0	.0	60	.0	.0	60	.0	.0
N10	25	.01	.08	25	.0	.0	25	.01	.09	25	.0	.0
N11	25	.0	.0	25	.0	.0	50	.0	.0	30	.0	.0
N12	30	.0	.0	25	.0	.0	25	.0	.0	25	.0	.0
N13	50	.0	.03	90	1.0	.0	60	.01	.11	40	1.0	.04

Note.  $N_R$  = number of essays rated.  $x$  Est. = rater halo parameter estimate.  
<sup>a</sup> “Halo raters” are raters with  $x$  estimates  $> .50$  in Exam 1. <sup>b</sup> “Non-halo raters” are raters with  $x$  estimates  $\leq .50$  in Exam 1 and scoring essays in the other three exams.

performance in terms of the ranks achieved, the potential impact of even slight differences in proficiency estimates under each model becomes readily apparent. For this purpose, we compared the RFM-based examinee rank order to the rank order derived from the MRFM-H proficiency estimates. This comparison yielded non-negligible absolute rank-order differences, as summarized in Table 5.

For example, in Exam 1, the mean absolute rank change (MARC) was 16.32 ( $SD = 14.94$ ); examinees’ rank orderings differed on average by about 16 ranks, depending on which model was used for estimating their proficiency.

However, compared to the other exams, the greatest MARC was obtained for Exam 3; the absolute rank-order difference ranged from 0 to 141 ( $M = 10.37$ ,  $SD = 10.15$ ). Considering the relatively small examinee sample size in Exam 3 ( $N = 2,619$ ), this MARC value is particularly striking. Figure 2 shows the frequency distribution of rank-order differences. This distribution was highly positively skewed. For the other exams, the skewness was somewhat smaller (Table 5).

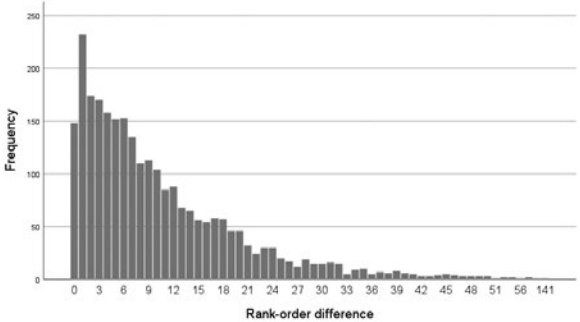
Rank differences of the magnitude observed here may have critical consequences for individual examinees. For example, consider

**Table 5**  
*Examinee Rank-Difference Statistics Based on RFM and MRFM-H Proficiency Estimates in Four Exams*

Statistic	Exam 1	Exam 2	Exam 3	Exam 4
Examinees ( $N$ )	5,191	5,501	2,619	2,366
Range	106	121	141	57
Range%	2.04	2.20	5.38	2.41
Median	12	14	7	6
90th percentile	37	44	23	19
Mean	16.32	19.33	10.37	8.23
$SD$	14.94	18.06	10.15	7.77
Skewness	1.47	1.56	2.26	1.66

Note. Rank-difference statistics refer to absolute differences. Range = maximum absolute rank change. Range% = maximum absolute rank change proportion.

**Figure 2**  
*Frequency Distribution of Rank Differences Between MRFM-H and RFM Examinee Proficiency Estimates (Exam 3)*



two scenarios where selection decisions are made based on the assessment outcomes. In the first scenario, an examination board adopts a pre-select strategy to decide the number of examinees to admit to a popular field of study. According to this strategy, the top 200 of 5,191 examinees (Exam 1) are selected. Then, among the most qualified examinees, six would be disregarded, given that their proficiency was estimated based on the RFM instead of the MRFM-H.

In the second scenario, the board adopts a pre-reject strategy to decide the number of examinees to reject. The bottom 100 of 5,501 examinees (Exam 2) are not considered further. Then, among the least qualified examinees,

seven would still have the chance to be selected based on the RFM instead of the MRFM-H proficiency estimates. Table 6 presents these and other examples of the possible impact of illusory halo effects under the pre-select and pre-reject decision strategies separately for each exam.

**Summary and Discussion**

There is a long tradition of research on the rating quality in performance assessments documenting various rater effects, errors, and biases (Eckes, 2023a; Engelhard & Wind, 2018; Guilford, 1936; Knoch et al., 2021; Myford & Wolfe, 2003). The research focus has been on the four “classic” errors, comprising severity/

**Table 6**

*Illustration of Illusory Halo Effects' Impact on Pre-Select and Pre-Reject Decisions Based on MRFM-H vs. RFM Examinee Rank Orders in Four Exams*

Decision strategy	Exam 1	Exam 2	Exam 3	Exam 4
Pre-select				
Top-100	1	1	1	1
Top-200	6	2	3	2
Top-300	2	7	2	1
Top-400	5	3	3	2
Top-500	4	8	5	3
Top-1000	7	10	7	3
Pre-reject				
Bottom-100	2	7	2	2
Bottom-200	5	3	4	0
Bottom-300	2	6	6	7
Bottom-400	4	9	6	5
Bottom-500	5	6	6	4
Bottom-1000	13	6	4	4

*Note.* In each exam, entries listed for pre-select decisions give the number of examinees ranked higher under the MRFM-H (the reference model) than under the RFM. Entries listed for pre-reject decisions give the number of examinees ranked lower under the MRFM-H than under the RFM.

leniency, halo, centrality/extremity, and restriction of range (Myford & Wolfe, 2003; Saal et al., 1980). According to Myford and Wolfe (2003), “the halo effect has been the most studied and has received the widest attention in the research literature” (p. 395). However, unlike the other classic rater effects, particularly severity/leniency and central tendency (Eckes, 2019, 2023b; Eckes & Jin, 2021a, 2021b; Engelhard & Wind, 2018), halo effects have proved difficult to pin down exactly (Lai et al., 2015; Murphy et al., 1993; Myford & Wolfe, 2004). The critical problem with halo effect detection is distinguishing true from illusory halo. In the present research, this problem is addressed using an extended many-facet Rasch measurement approach, the MRFM-H (Jin & Chiu, 2022).

The MRFM-H combines latent class analysis with conjoint measurement of examinee proficiency, rater severity, criterion difficulty, and possibly further latent dimensions (Jin & Wang, 2017; Sen & Cohen, 2019; von

Davier & Rost, 2016). Notably, this model 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, the MRFM-H estimates the likelihood that he or she belongs to the class of raters exhibiting illusory halo effects.

Like the standard RFM, the MRFM-H quantifies the extent to which raters exhibit differences in severity and corrects for their impact on the assessment outcomes. However, the MRFM-H goes beyond the RFM because it detects illusory halo effects (if any) and allows researchers to control these effects' influence on the ratings statistically. This model simultaneously corrects the scores raters assign to examinees for severity and halo effects.

We used the MRFM-H to analyze the ratings of examinee performances on writing tasks presented in four successive assessments in the context of university admissions. The number of examinees ranged from 5,501 in Exam 2 to 2,366 in Exam 4, totaling over

15,600 examinees; the rater sample sizes ranged from 52 raters (Exam 4) to 107 (Exam 2), totaling over 160 raters. For comparison purposes, in Exam 1, we also computed two popular halo statistics based on the raw scores, the mean correlation between criteria ( $r_H$ ) and the mean within-examinee standard deviation across criteria ( $SD_H$ ). We also ran standard RFM analyses on the ratings.

### Answers to the Research Questions

We conducted the MRFM-H and RFM analyses adopting Bayesian MCMC procedures for parameter estimation (Gelman et al., 2013; Kruschke, 2015; Lunn et al., 2013). Commonly used data-model fit statistics (PSRF, PPP-value) attested to a satisfactory fit of the rating data to the MRFM-H and the RFM. Bayesian comparison indices for evaluating each model's relative fit to the data (DIC, WAIC, and LOOIC) consistently favored the MRFM-H over the RFM. Thus, the MRFM-H outperformed the RFM in each exam, answering the first research question.

The evidence attesting to illusory halo effects in our rater samples gave rise to a closer look at the raters' halo tendencies in the first exam, identifying classes of halo and non-halo raters and tracking their rating behavior across the other three exams. Regarding the second research question, in Exam 1, we found 11 raters (out of a total of 100 raters) exhibiting illusory halo effects. One of these raters qualified as a borderline case since the halo parameter estimate ( $x_j$ ) was only slightly above the cut-off value of .50. Among the 89 non-halo raters, we identified 13 raters who also participated in Exams 2 through 4, forming the comparison sample for the halo raters across exams.

Research Question 3 addressed the issue of stability and change in raters' latent class membership across the four exams. The first exam was administered in November 2019, and the last was in September 2020, thus spanning 10 months. Among the 11 raters belonging to the halo class, 10 raters participated in at least

two exams. Three of these raters consistently showed halo tendencies in all exams in which they participated. Another six raters changed from halo to non-halo in the last exam in which they participated; two had consistently exhibited halo tendencies in the first three exams. One rater was identified as a borderline case, showing slight halo tendencies in the first exam but switching to non-halo for the other two exams. In contrast to this mix of stability and change among the Exam 1 halo raters, twelve of the thirteen raters forming the comparison sample remained stable in their non-halo class membership over the other three exams. However, one non-halo rater exhibited erratic rating behavior, changing his or her class membership every other exam.

Standing in the observed ratings research tradition (Wind & Peterson, 2018), previous proposals for halo detection included computing criterion intercorrelations (sometimes in conjunction with a factor or principal-component analysis), within-examinee variability statistics, and examinee-rater interactions (Saal et al., 1980). For comparison purposes, we computed the two most popular raw-score halo statistics, the mean correlation between the scoring criteria,  $r_H$ , and the mean within-examinee standard deviation across criteria,  $SD_H$  (Fiscaro & Vance, 1994; Pulakos et al., 1986). Using the halo rater versus non-halo rater classification resulting from the MRFM-H analysis of Exam 1 ratings as a reference, we found that neither the  $r_H$  statistic nor the  $SD_H$  statistic reliably distinguished between these two classes of raters. This finding is reminiscent of the reservations Fiscaro and Vance (1994) expressed regarding using these two halo statistics.

The fifth and final research question dealt with the practical consequences of illusory halo effects for examinees in the context of high-stakes assessments. In the assessments considered here, the writing scores informed decisions about examinees' admission to institutions of higher education in Germany. As mentioned, about 10% of the raters were subject

to illusory halo across exams—a non-negligible percentage. Probing more deeply into the halo effects' impact on the assessment outcomes, we computed the absolute differences between the examinee rank orders resulting from the standard RFM and the MRFM-H proficiency estimates. Relative to the examinee sample size, the most significant number of rank changes occurred in Exam 3; the maximum absolute rank change proportion was 5.38%, with a maximum rank change of 141, depending on whether the RFM or the MRFM-H was used to estimate the examinees' proficiency.

The differences between examinee rank orders under the RFM and the MRFM-H across the four exams may influence high-stakes decision-making, such as selection decisions in university admissions (Table 6). Considering the examinee sample sizes ranging from 2,366 (Exam 4) to 5,501 (Exam 2), absolute rank changes in the order of 10 places or less may seem insignificant from a purely statistical perspective. However, they can have severe consequences for individual examinees' study and life plans. Therefore, accounting for illusory halo effects in raters' judgments of examinee performances helps increase an assessment's validity and fairness.

### Applied and Basic Research Perspectives

The information gained from a mixture Rasch facets analysis can be used to improve rater training, monitoring, and selection. For example, providing raters with individualized feedback on their halo tendencies, possibly along with suitably prepared charts or diagrams depicting their severity or leniency (e.g., Hoskens & Wilson, 2001; Huang & Chen, 2022; Stahl & Lunz, 1996), can raise their awareness of faulty rating patterns some of which may have become habitual over time. Creating awareness is a prerequisite for successfully implementing training programs to correct rater misbehavior (Knoch et al., 2021). Also, continuous rater monitoring focusing on their rating tendencies across two or more assessments helps control the long-term effects

of training and related intervention measures. For example, in the present study, some halo raters changed their rating tendency across exams, and they did so in the desired non-halo direction (Table 4). Follow-up studies could examine how robust such a change is.

A similarly important use of halo information based on MRFM-H findings concerns selecting raters for operational rating sessions. For example, raters repeatedly exhibiting illusory halo effects or changing their rating tendency in unpredictable ways across exams may be considered for exclusion from the panel of operational raters. Alternatively, exam bodies may recommend that these raters undergo retraining or recalibration to get closely aligned again with the assessment goals and the scoring guidelines. In the present analysis, non-halo Raters N01 to N12 (Table 4) were functioning as intended, qualifying as candidates for continued participation in operational scoring sessions. By contrast, Rater N13's erratic behavior at least signals an urgent need for retraining.

Even in assessment contexts typically not considered high-stakes, knowing which raters tend to exhibit illusory halo can become relevant. Thus, in diagnostic assessments, where examinees (or learners) expect to get detailed information on their strengths and weaknesses across distinct performance dimensions (Lee, 2015), the participation of halo raters diminishes the informative value of learner feedback. To the extent that halo tendencies prevail in the rater group, the learning profiles will flatten and lose their function as a valuable tool for promoting learning progress.

In rating quality studies, a critical distinction refers to the frame of reference for evaluating a rater's performance (Myford & Wolfe, 2009; Wolfe, 2020). In the present research, we adopted an internal frame of reference, assuming an equal status of raters in the group and examining rater behavior "in terms of the degree to which the ratings of a particular rater agree with the ratings that other raters assign" (Myford & Wolfe, 2009, p. 372).

In other words, the target against which we compared the scores a particular rater assigned were the other raters in Exam 1, leading to distinguishing between the halo and non-halo raters (the same applied to Exam 2 to Exam 4 raters).

This view contrasts with an external frame of reference, examining rater behavior "in terms of the degree to which the ratings of a particular rater agree with scores on an external criterion" (Myford & Wolfe, 2009, p. 373). Typically, the external criterion is a set of ratings assumed to be valid or "true" obtained from a single expert or an expert group. Under the external frame of reference perspective, a mixture Rasch facets analysis helps ensure the rating quality in at least two ways. First, MRFM-H parameter estimates allow researchers to select those raters for the role of experts who are demonstrably not subject to halo, severity, and possibly other rating tendencies (e.g., centrality). Second, monitoring expert raters' rating tendencies over an extended period can help evaluate how close they come to the ideal of providing error-free scores.

Another promising line of research concerns broadening the perspective on the cognitive or judgmental processes involved in bringing about illusory halo effects. Thus, besides inadequate criterion discrimination, Fiscaro and Lance (1990) discussed two other psychological mechanisms: general impressions and a single, salient performance feature impacting the ratings. These mechanisms may become a worthwhile goal of future research. Similarly, studies of illusory halo could combine Rasch-based measurement models like the MRFM-H with more qualitatively oriented approaches based on structured interviews, stimulated recall, or verbal protocol analysis (Myford, 2012; Turner, 2014).

From a psychometric modeling perspective, further extensions of the MRFM-H may be considered (Jin & Chiu, 2022). For example, in its current version, this model incorporates two rater parameters, representing severity and halo effects. However, other rater parameters

may also be included. Notably, Jin and Wang (2018) proposed a Rasch facets model incorporating a rater centrality parameter, the "facets model–severity and centrality" (FM-SC). Therefore, building a more general facets model combining the MRFM-H and the FM-SC, simultaneously addressing rater severity, centrality, and illusory halo effects under a Bayesian estimation framework, promises to yield a more comprehensive picture of the raters' performance.

### Limitations

The findings from our longitudinal writing assessment study provided evidence of the MRFM-H's efficiency and utility for detecting and measuring illusory halo effects. However, several limitations of the present research should also be noted.

We analyzed rating data from four writing exams extending over 10 months. Such a time interval may be too short to provide decisive information on the issue of stability and change of halo tendencies. Previous research using a longitudinal design has focused on changes in rater severity, examining time intervals ranging from a few hours, days, or weeks (Congdon & McQueen, 2000; Lunz & Stahl, 1990; Wilson & Case, 2000) to 10 or more years (Lamprianou et al., 2021; Neittaanmäki & Lamprianou, 2024a, 2024b; O'Neill & Lunz, 2000). Therefore, the present period was somewhere in between but must be extended in future research.

Another critical issue concerns the performance links design used in the four exams. For example, in Exam 1, each rater scored 55 performances on average, including the three performances (the "links") common to all raters' workloads. As a result, the proportion of missing ratings in this exam (and similarly in the other exams) was very high (over 90%), lowering the precision of parameter estimation. As Jin and Chiu (2022) concluded in their simulation study, comparing performance links and systematic links designs, having more raters score the same set of performances raises the measurement precision, which, in turn, leads to

higher rater classification accuracy. At least, the present rating design was sufficient to detect the halo tendencies of 11 Exam 1 raters.

Jin and Chiu (2022) discussed two more factors with direct implications for the accuracy of rater halo classification: (a) the number of scoring criteria and (b) the magnitude of the differences between criterion difficulty estimates (i.e., psychometric criterion heterogeneity). The theoretical structure underlying the MRFM-H suggests that the accuracy of classifying raters as illusory halo or non-halo raters will be higher when using more scoring criteria, and the criteria have significantly different difficulty measures. Our findings support at least the first part of this reasoning. Ratings were provided on a relatively large set of nine criteria, unlike the Eckes and Jin (2022) study, where raters used only three criteria. In contrast to that study, we found significant halo effects in the present research. However, the set of nine criteria showed a relatively small range of difficulties. Of course, definite conclusions regarding the factors influencing MRFM-H classification accuracy call for further studies, notably simulation studies systematically varying the levels of these two factors.

Finally, due to the COVID-19 pandemic starting to impact Germany in the spring of 2020, the number of examinees taking Exams 3 and 4, respectively, was roughly halved compared to the previous exams. The same holds for the number of operational raters. These unforeseen circumstances limited the generality of the present findings. Thus, the number of raters whose halo tendencies could be tracked across all four exams dropped significantly. Nonetheless, the consistency of the Bayesian model fit comparisons favoring the MRFM-H over the RFM across the varying examinee and rater samples provides a firm base for our main conclusions regarding the detection and measurement of raters' halo tendencies.

### Conclusions

Our longitudinal study of halo effects was

situated in a high-stakes writing assessment context. Using the mixture Rasch facets model for halo effects (MRFM-H; Jin & Chiu, 2022), we separated true from illusory halo effects in four writing exams administered across 10 months. In each exam, the halo model parameter estimates identified raters subject to illusory halo (halo raters) and raters free from halo tendencies (non-halo raters). On average, we found that about 10% of the operational raters belonged to the halo rater class. These percentages are non-negligible and must be considered when evaluating the ratings' psychometric quality. Furthermore, the halo tendencies some raters exhibited had a demonstrable impact on making selection decisions on examinees. As a member of the many-facet Rasch model family, the MRFM-H corrects for the impact of raters' illusory halo tendencies on each examinee's final score, adding to the validity and fairness of the score's interpretation and use.

Tracking 11 halo and 13 non-halo raters identified in the Exam 1 analysis across the other three exams, we found evidence for stability and change among halo raters, with some raters shifting from halo to non-halo and high stability among the non-halo raters. Thus, most raters were non-halo throughout the exams or adhered to their non-halo tendencies, which developed over time. Notably, traditional raw-score halo statistics failed to distinguish between these two latent classes of raters in Exam 1, rendering these statistics useless in exploring halo tendencies across the other exams. In contrast, the MRFM-H reliably detected raters subject to illusory halo effects and thus proved a valuable tool for examining such effects in performance assessments.

### References

- AlHakmani, R., & Sheng, Y. (2023). Empirical evaluation of fully Bayesian information criteria for mixture IRT models using NUTS. *Behaviormetrika*, 50, 93–120. <https://doi.org/10.1007/s41237-022-00167-x>
- Baghaei, P., & Carstensen, C. H. (2013).

Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research, and Evaluation*, 18(5), 1–13. <https://doi.org/10.7275/n191-pt86>

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>

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>

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

DeMars, C. E., Shapovalov, Y. A., & Hathcoat, J. D. (2023, April 13–15). *Many-facet Rasch designs: How should raters be assigned to examinees?* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Chicago, IL, United States. <https://commons.lib.jmu.edu/cgi/viewcontent.cgi?article=1071&context=gradpsych>

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. (2023a). Detecting and measuring rater effects in performance assessments: Advances in many-facet Rasch modeling. In N. Dobrić, H. Cesnik, & C. Harsch (Eds.), *Festschrift in honour of Günther Sigott: Advanced methods in language testing* (pp. 195–223). Peter Lang. <https://doi.org/10.3726/b21019>

Eckes, T. (2023b). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang. <https://doi.org/10.3726/b20875>

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](https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2021/Seiten_aus_PTAM_2021-1_ebook_4.pdf)

Eckes, T., & Jin, K.-Y. (2022). Detecting illusory halo effects in rater-mediated assessment: A mixture Rasch facets modeling approach. *Psychological Test and Assessment Modeling*, 64(1), 87–111. [https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam\\_2022-1/PTAM\\_1-2022\\_5\\_kor.pdf](https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_2022-1/PTAM_1-2022_5_kor.pdf)

Engelhard, G. (1994). Examining rater errors in the assessment of written composition

- with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93–112. <https://doi.org/10.1111/j.1745-3984.1994.tb00436.x>
- Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge. <https://doi.org/10.4324/9781315766829>
- 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), 366–371. <https://doi.org/10.1177/0013164494054002010>
- 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/de/ueber-testdaf/testdaf-ein-produkt-der-gast/informationsmaterialien/>
- 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](https://projecteuclid.org/download/pdf_1/euclid.ss/1177011136)
- Guilford, J. P. (1936). *Psychometric methods*. McGraw-Hill.
- 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>
- Huang, J., & Chen, G. (2022). Individualized feedback to raters in language assessment: Impacts on rater effects. *Assessing Writing*, 52, 100623. <https://doi.org/10.1016/j.asw.2022.100623>
- Jin, K.-Y., & Chiu, M. M. (2022). A mixture Rasch facets model for rater's illusory halo effects. *Behavior Research Methods*, 54(6), 2750–2764. <https://doi.org/10.3758/s13428-021-01721-3>
- 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>
- Kim, H. (2020). Effects of rating criteria order on the halo effect in L2 writing assessment: A many-facet Rasch measurement analysis. *Language Testing in Asia*, 10, Article 16. <https://doi.org/10.1186/s40468-020-00115-0>
- Knoch, U., Fairbairn, J., & Jin, Y. (2021). *Scoring second language spoken and written performance: Issues, options and directions*. Equinox.
- Kreitchmann, R. S., de la Torre, J., Sorrel, M. A., Nájera, P., & Abad, F. J. (2023). Improving reliability estimation in cognitive diagnosis modeling. *Behavior Research Methods*, 55(6), 3446–3460. <https://doi.org/10.3758/s13428-022-01967-5>
- Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press/Elsevier.
- 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>
- 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>
- Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing*, 32(3), 299–316. <https://doi.org/10.1177/0265532214565387>
- Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Chapman & Hall/CRC.
- Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28(4), 543–560. <https://doi.org/10.1177/0265532211406422>
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. 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, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, 13(4), 425–444. <https://doi.org/10.1177/016327879001300405>
- Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibrations. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 99–112). Ablex.
- 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/fileadmin/download/ptam/2-2017\\_20170627/03\\_Luo\\_.pdf](https://www.psychologie-aktuell.com/fileadmin/download/ptam/2-2017_20170627/03_Luo_.pdf)
- 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.
- Murphy, K. R. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology*, 67(2), 161–164. <https://doi.org/10.1037/0021-9010.67.2.161>
- 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>
- 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>
- Neittaanmäki, R., & Lamprianou, I. (2024a). All types of experience are equal, but some are more equal: The effect of different types of experience on rater severity and rater consistency. *Language Testing*, 41(3), 606–626. <https://doi.org/10.1177/02655322241239362>
- Neittaanmäki, R., & Lamprianou, I. (2024b). Communal factors in rater severity and consistency over time in high-stakes oral assessment. *Language Testing*, 41(3), 584–605. <https://doi.org/10.1177/02655322241239363>
- Norris, J., & Drackert, A. (2018). Test review: TestDaF. *Language Testing*, 35(1), 149–157. <https://doi.org/10.1177/0265532217715848>
- O’Grady, S. (2023). Halo effects in rating data: Assessing speech fluency. *Research Methods in Applied Linguistics*, 2(2), 100048. <https://doi.org/10.1016/j.rmal.2023.100048>
- O’Neill, T. R., & Lunz, M. E. (2000). A method to study rater severity across several administrations. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 135–146). Ablex.
- Plummer, M. (2017). *JAGS version 4.3.0 user manual*. [https://people.stat.sc.edu/hansont/stat740/jags\\_user\\_manual.pdf](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. (2022). *R: A language and environment for computing* (Version 4.2.2) [Computer software]. R Foundation for Statistical Computing. <http://www.R-project.org>
- Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14(3), 271–282. <https://doi.org/10.1177/014662169001400305>
- Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 25–42). Springer.
- 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>
- Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, 17(4), 177–191. <https://doi.org/10.1080/15366367.2019.1583506>
- 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>
- 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>
- Stahl, J. A., & Lunz, M. E. (1996). Judge performance reports: Media and message. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 113–125). Ablex.
- 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.
- Uto, M. (2023). A Bayesian many-facet Rasch model with Markov modeling for rater severity drift. *Behavior Research Methods*, 55(7), 3910–3928. <https://doi.org/10.3758/s13428-022-01997-z>
- 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). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models* (Version 2.4.1) [Computer software]. <https://cran.r-project.org/web/packages/loo/index.html>
- von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 393–406). Chapman & Hall/CRC.
- 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>
- Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (pp. 113–133). Ablex.
- 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., & 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. (2020). Human scoring with automated scoring in mind. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 49–67). Chapman & Hall/CRC. <https://doi.org/10.1201/9781351264808>
- 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.