Applied Research on English Language

V. 12 N. 1 2023
pp: 159-176
http://jare.ui.ac.ir

DOI: 10.22108/ARE.2022.132503.1848
Document Type: Research Paper

Detecting Halo Effects across Rubric Criteria in L2 Writing Assessment: A Many-facet Rasch Analysis

Hakimeh Ayoobiyan [1*], **Alireza Ahmadi** [2]

[1] *PhD Candidate, Department of Foreign Languages and Linguistics, Faculty of Humanities, Shiraz University, Shiraz, Iran*
[2] *Professor, Department of Foreign Languages and Linguistics, Faculty of Humanities, Shiraz University, Shiraz, Iran*

Received: 2022/01/28 **Accepted: 2022/07/11**

Abstract: This study applied multi-faceted Rasch measurement to investigate the halo effect in the performance-based assessment of writing across rubric criteria. Five raters who had received specialized training applied a four-criteria rating rubric to analytically rate writing scripts on two argumentative topics. Facets, a Rasch computer program, was utilized to pinpoint the halo effect by analyzing rater and rubric criteria interactions. After examining the appropriateness of the rubric in terms of functionality, the results showed that except for one rater, the raters did not exhibit any sign of the halo effect across rubric criteria. Generally, the severity hierarchies for raters and difficulty levels for rubric criteria suggested that raters' training and their perceptions of the importance of criteria were associated with their manifestation of the halo effect. Pedagogically, through a detailed facet analysis of interactions between raters and rubric criteria, rater trainers may better realize how to design effective training programs to minimize raters' variance including the halo effect and improve the overall objectivity of human rating.

Keywords: Halo Effect, Rubric Criteria, Rater Variance, Many-facet Rasch Measurement (MFRM).

* **Corresponding Author.**

Authors' Email Address:
[1] **Hakimeh Ayoobiyan** (s.ayoobiyan@gmail.com), [2] **Alireza Ahmadi** (arahmadi@shiraz.ac.ir)

Introduction

In assessing performance-based skills such as writing and speaking, examinees are required to demonstrate their ability on a number of assignments, and the quality of their response is commonly assessed via human judgment. This simple fact may result in a myriad of complications, the most noticeable one being that judges will usually disagree. Although raters are expected to strictly follow rating scales without imposing any bias, it seems that even experienced or well-trained raters are unable to avoid bias (Brown, 1995; Brown, 2003; Cohen, 1994; Cumming, Kantor, & Powers, 2002; Hamp-Lyons & Davies, 2008; Jeong; 2017; Lumley, 2006). Eckes (2012) believed while an acceptable level of agreement may be observed within raters and between raters, variations in raters' bias might still exist. Since raters' bias may affect rating quality and fairness, it is crucial to investigate what may lead raters to show bias.

The halo effect as a source of raters' cognitive bias occurs when judgments of a certain rating criterion are influenced by that of other rating criteria in a positive or negative way (Anderson, 2015; McDonald, 1999). The halo effect manifests itself as the erroneously inflated correlations among distinct criteria in an analytic rating rubric. As raters are subject to the halo effect, they tend to assign more similar or even identical scores across rating criteria in a rubric than they should. Consequently, the halo effect restricts the value of diagnostic feedback given to the test takers and neutralizes the attempts in developing multiple rating scales. The sources of the halo effect are proposed as raters' general impression, a salient rating impression, and an inability of raters induced by insufficient training (Ballard, 2017; Lance, LaPointe, & Stewart, 1994).

There is a growing consensus in the literature in which researchers could examine the halo effect by running Multi-Faceted Rasch Measurement (MFRM). MFRM considers raters as an individual facet, and this provides the opportunity for researchers to examine possible rating variance by examining the interactions of raters with other facets (Linacre, 2018a). To provide in-depth and accurate information on examinees' writing proficiency, MFRM has the potential to fulfill the function of identifying the exact point where rating variance occurs (Andrich & Marais, 2019; Weigle, 1998).

Although previous studies have shed some light on raters' halo effect in the context of performance-based assessment (Andrich, Humphry, & Marais, 2012; Andrich & Kreiner, 2010; Bechger, Maris, & Hsiao, 2010; Engelhard Jr, 1994; Lai, Wolfe, & Vickers, 2015;

Myford & Wolfe, 2003, 2004), what remains blurred is whether experienced and trained raters show the halo effect across criteria difficulty. To address the call for continued research on the halo effect, the present study is going to examine the extent to which the halo effect may be detected across criteria difficulty of the rubric in trained raters' ratings using MFRM analysis.

Literature Review

MFRM

MFRM rooted in Item Response Theory (IRT) is an extension of the basic, dichotomous Rasch model. The Rasch model is a one-parameter model in IRT in which only item difficulty is defined as a parameter. In this model, the probability of giving a correct response by a test taker is estimated by measuring the difference between a test taker's ability and item difficulty (Rasch, 1980). Compared to Rasch, MFRM is a linear model logistically transforming polytomous scores on a performance test into an interval logit scale (Eckes, 2015). In MFRM, analyses are done by first forming hypotheses about the facets of an assessment, specifying a model that includes all the relevant facets, and finally calibrating the facets using the observed scores (Fan & Bond, 2019; Lynch & McNamara, 1998).

Three assumptions require to be met in MFRM: unidimensionality, local independence, and certainty of response. Unidimensionality means that score differences are attributed to one single trait. Local independence refers to the fact that responses to each test item and task are independent of one another. The certainty of responses is the test-takers exert efforts to complete tasks (Ockey, 2012). It is usually difficult to meet all of these assumptions in language testing if any (Ockey, 2012). The difference between this measurement model and G-theory is that MFRM does not require the assumption of random sampling because MFRM builds on a scaling theory (Barkaoui, 2014; Brennan, 2001; Kim & Wilson, 2009). Scaling theory is "a branch of measurement theory that focuses on rationales and mathematical techniques for determining what numbers should be used to represent different amounts of the property being measured" (Allen & Yen, 2001, p. 179).

As considered for research designs, the main assumption for MFRM is a logical connection between disjoined sets of observations (Kim & Wilson, 2009). Furthermore, to obtain accurate parameter estimates, MFRM requires a sufficiently large sample in which at least 100 observations should be included (Ockey, 2012). If parameter estimates are not

precise due to the small sample size, the standard errors of the parameter become larger; consequently, the statistical power of the fit statistics decreases (Barkaoui, 2014).

One of the important features of MFRM is that it provides individual-level information, which is very helpful for inspecting the function of each of the elements in each facet (Barkaoui, 2014; Eckes, 2015; Linacre, 2018b). But the major advantage of MFRM is measurement invariance (Eckes, 2015) which is defined as while data fit the identified model, examinees' ability is constant across test items, and item difficulty is constant regardless of test-taker groups. Because of this significant feature, test scores are regarded as enough statistics for estimating a test taker's ability in terms of reliability (Eckes, 2015).

Testing researchers commonly use MFRM analysis to examine different effects of various facets while controlling for the effects of other facets. For example, if a test taker's performance on an argumentative writing task is rated by a relatively severe rater, the test taker's ability estimate is adjusted, controlling for the rater's severity. Besides, diagnostic information from MFRM can be used for various purposes (Andrich & Marais, 2019). For instance, for those raters whose ratings do not fit the pattern, rater training for the raters can be provided (Lynch & McNamara, 1998; Sudweeks, Reeve, & Bradshaw, 2004). Similarly, the diagnostic information about individual tasks and a rating rubric can be used for the revision of the tasks and rating rubric (Lynch & McNamara, 1998). MFRM is related to bias analysis which functions similarly to differential item functioning (DIF), and permits the identification of a specific combination of elements across facets (Barkaoui, 2014; Sudweeks et al., 2004). For example, bias analysis identifies if individual raters rate the performance of a group of test-takers differently from the other test-takers. FACETS (Linacre, 2016) as a Rasch computer program is utilized to provide information such as standard errors and fit statistics to detect rater variance (Eckes, 2015).

Rater Variance in Performance-based Assessment

One of the main concerns of test constructors when assessing writing through 'direct' assessment is scoring validity in general and rater variance in particular. Scoring validity includes all the aspects that may block test scores' overall validity in the rating process, such as rating criteria, procedures, conditions, rater training, and awarding (Weir, 2005). Rater variance is referred to the variance in scores assigned by raters on the same script, using the same rating scale (McNamara, 2000). Rater variance as the potential "Achilles heel of

performance testing" (O'Sullivan & Rignall, 2007, p. 47) is considered a significant source of construct-irrelevant variance.

There are a number of sources of rater variance that may influence scores, and they are of great concern to language testers (McNamara, 1996; Lumley, 2006; Vaughan, 1991; Weigle, 2002). So, in writing assessment, a test taker's score may be due not only to the test taker's writing ability but also to the rater scoring process of the script. For instance, another rater could very well award a completely different score to the same written script which is deemed a threat to scorer validity. McNamara (1996, 2000) and Myford and Wolfe (2003, 2004) focus on a number of ways in which raters may systematically differ:

• Raters may differ in their overall severity when scoring and one rater may consistently assign higher marks than another rater on all scripts.

• Raters may interact differently with a specific item or type of candidate which is called the bias effect. They may result in being consistently more lenient to one type of item/test taker(s) and severe on another type of item/test taker(s).

• Raters may have different interpretations of the scoring criteria or the descriptors on the rating scale.

• Raters' assigned scores on distinct criteria (especially in the case of the analytic scale) may be influenced by their overall impression of the performance (halo effect).

• Raters may avoid extreme ends in a rating scale and assign scores closer to the mid-point (central tendency effect).

• Raters may not be self-consistent. The same rater may award a different score to the same script when scored more than once in a different context (e.g., time of day, order of compositions, single or group rating), based on their expectation of the script (McNamara, 2000; Weigle, 2002; Zhang & Lu, 2022) or different physical or emotional state. Myford and Wolfe (2003) called this phenomenon 'randomness'.

In addition to all the factors aforementioned, different physical (age, sex, illness, disabilities, etc.), psychological (personality, memory, motivation, cognitive style, etc.), and experiential characteristics (language background, education, experience, etc.) may influence raters' performance and result in rater variance (Li & Huang, 2022; Shaw & Weir, 2007). Shaw and Weir argue that the physical and psychological factors "may not lend themselves to future investigation or not be considered worth the effort" (ibid: p. 168), but experiential factors can be more easily recognized and addressed. In addition, experiential factors are more likely to have a systematic effect on raters. Conventionally, rater variance has been

viewed as a problem and testers generally want a stronger agreement amongst raters in order to reduce differences as much as possible, i.e., a higher reliability coefficient (Weigle, 1998). In an attempt to account for the rater variance, many administrators of performance-based writing assessments have opted to employ specialized rater training.

Halo Effects

For the first time, the term halo effect was coined by Thorndike (1920) and defined it as a marked tendency to think of the person in general as either good or inferior and to color the judgments of the qualities by their general impression. In particular, he characterized the halo effect as "suffusing ratings of special features with a halo belonging to the individual as a whole" (p. 25). But in the field of language testing, it was first investigated by Yorozuya and Oller Jr (1980) and they named it "judgmental bias" and defined it as "a tendency for judges to assign similar scores across the various scales". They believed that this kind of judgment could be called a halo effect.

In the context of MFRM analysis, the halo effect manifests itself as raters' tendency to award similar ratings on conceptually distinct constructs (Myford & Wolf, 2004). In the light of MFRM, Eckes (2009) rightly referred to the fact that the halo effect as a source of rater variance may be detected when raters fail to distinguish between distinct features of test takers' performance but rather assign similar ratings across those features.

Halo effects can occur for many reasons. For instance, judges may formulate a general impression after having seen a few performances, and following judgments may be to a great extent influenced by the first impression. Some raters may simply stop paying attention to the examinees' performances whereas others may (unconsciously) be tempted to make subsequent ratings consistent with earlier ratings. Halo effects may also be due to contrast effects: that is, the standard of a previously rated student may influence the ratings. Just in a case that the data have been gathered in a carefully designed experiment, it is only possible to claim about the nature and possible causes of halo effects, for little is known about the complex cognitive processes of human scoring (e.g., Lumley, 2006; Vaughan, 1991). This is why it is believed that it is useful to have a method to detect halo effects that do not require detailed assumptions about rater behavior.

Halo effects imply a decrease in the number of independent opportunities for the candidate to demonstrate their proficiency and diminish the reliability of the test. At the same time, the correlation between ratings of the same candidate may be increased, which gives the

impression that the test is very reliable. In the extreme case, only the first performance is rated and all subsequent ratings are equal to the first, which leads to high correlations between scores on different parts of the examination (Marais & Andrich, 2011). However, because only one performance was rated it cannot be expected to make very precise statements about the examinee's ability.

In the field of second language learning, there are a couple of studies that have examined the halo effect employing MFRM. Engelhard Jr (1994) recruited 15 highly experienced, trained raters to rate writing scripts, using an analytic scale with five criteria. He focused on four aspects of raters: severity or leniency, halo, central tendency, and restriction of range. The results of his study showed that two out of 15 raters' ratings manifested the halo effect, which means that they tended to rate holistically and failed to differentiate among students.

In another study, Kozaki (2004) employed facets to analyze raters' behavior. To that end, four professionally and bilingual raters were chosen to judge the performance of the examinees on a 4-point rating scale in seven rating categories. Raters' rating was analyzed by facets to examine their severity, central tendency, and halo effect. She found that two of the raters assigned unexpectedly harsh ratings to the test taker with the lowest level of ability and unexpectedly lenient ratings to the test taker with the highest level of ability. She interpreted such unexpected ratings as signs of the halo effect, in which "judges carry over the impression of competence… creating non-independence of assessment categories… grammar or vocabulary or both" (Kozaki, 2004, pp. 21-22).

In a more recent study, Knoch, Read, and Von Randow (2007) compared the effectiveness of online and face-to-face feedback to individual raters within the context of a large-scale academic writing assessment of students. Sixteen highly trained experienced native and non-native English teachers were equally divided into online and face-to-face groups. The raters used a three-criteria rubric to rate 70 candidates' writing scripts. Four rater effects of rater severity, internal consistency, central tendency, and halo effect were examined. Using group and individual statistics indicators of facets, the authors claimed that, at the group level, there was no sign of the halo effect but at the individual level, there was very low rater fit mean square indices which could be an indication of the halo effect.

In the latest study conducted on the halo effect, it was found that the rating criteria played a crucial role on average received much harsher or lower ratings from the raters (Lai

et al., 2015). Their findings indicated that organization as the criterion included in the rubric was most subject to the halo effect.

The Present Study

While the halo effect has been identified as a source of raters' construct irrelevant variance in the performance-based assessment such as writing, it is obvious that the magnitude of the halo effect in relation to criteria difficulty has been untouched. In order to redress this gap, the present study uses many-facet Rasch measurements to detect the magnitude of the halo effect exhibited by trained raters across criteria difficulty. So, the main objective of this study is to answer the following research question:

To what extent do highly trained and experienced raters display the halo effect across rubric criteria when rating the writing scripts?

Method

Participants

The participants were five experienced male raters recruited to take part in this study. They had different rating experiences ranging from 8 to 21 with a mean of 11.22 years. All raters had Bachelor's degrees in a relevant discipline (English Literature and Translation) or higher qualifications at the time of the study with two holding Ph.D. degrees (in English Literature and English Language Teaching). The raters were selected based on snowball sampling (Ary, Jacobs, Sorensen, & Walker, 2014). All five selected raters had rated IELTS writings tasks 1 and 2 for so many years and could be regarded as experienced raters. Each rater received training separately and participated in the specialized IDP (Individual Development Plan) programs which were held by IELTS centers. In this program, the rubrics on both tasks 1 and 2 were elaborated and raters received elaborated training on the rating process of both tasks by a native IELTS examiner.

Instruments and Materials

Rating Scale

IELTS writing task 2 rubric (the public version available at https://takeielts.britishcouncil.org /sites/default/files/ielts_task_2_writing_band_descriptors.pdf) is used for assessing writing scripts. The scale focuses on four aspects of writing (task response, coherence and cohesion, lexical resources, and grammatical range and accuracy), and consists of multiple scoring

criteria. The band scores are from 0 relating to the lowest possible performance to 9 indicating the highest possible performance on writing task 2. The qualitative descriptors are provided for each band score and each related criterion.

Writing Samples

A total of 80 argumentative writing scripts of participants in response to two tasks 2 from the IELTS exam were collected. To control for a possible topic-type effect (Hamp-Lyons, 1990), all students were given the same topic prompts to elicit samples of argumentative writing. The topics were selected in a way that did not require any special background knowledge on the part of students. Students had 40 minutes to spend on each task and write at least 250 words for completing each task.

Data Collection and Analysis Procedure

All five raters working in IELTS exam centers as experienced raters were recruited based on their availability at the time of the study. The assessors were briefed about the study and they were provided with 80 samples of writing and an accompanying public version of the rubric for each script. Each rater was provided with the prompt and the public version of the rubric. The rating samples were distributed to the raters so that each script was rated by at least two raters in the group of 40 students in each task 2. The raters each rated 16 essays for each task 2 and, in total, 32 essays were given to each rater to assess. Then, they were asked to rate each performance and they provided a whole score plus scores of each subscale for each script. They also were requested to highlight the descriptors for each of the criteria in the rubric. To ensure that there was sufficient overlap between writing scripts to allow for appropriate and meaningful statistical analysis, the rating plan was designed in a way that the highest connectivity among raters existed. Each script was rated by two raters, and all scripts were divided in a way that half of the ratings by each rater were shared by another rater. So, using the rubric, each essay was evaluated by two raters which provided for a suitable judging plan to use facet analysis (Linacre, 2010).

The rating data were subjected to a many-facet Rasch analysis using FACET (version 3.71.4, Linacre, 2016). A four-facet study was designed with the facets including the rater, test taker, task, and rubric. The Halo effect of each rater was investigated by means of rater fit statistics in relation to criteria difficulty and the patterns of each rater's ratings (Myford & Wolfe, 2004).

Results

For a rubric to function appropriately, evidence had to reveal that the various criteria in the rubric were combined to measure one underlying trait or multi-dimensionality that existed in the data. First, this had been checked through a correlation matrix of four criteria in the rubric. As shown in Table 1, the correlation ranged from 0.39 to 0.66, indicating some relationships but little indication that any of the criteria is redundant.

Table 1. *Subscale Correlations*

	TR	C&C	LR	GR&A
Task Achievement	*			
Coherence & Cohesion	0.56	*		
Lexical Resources	0.39	0.56	*	
Grammatical Range & Accuracy	0.41	0.57	0.66	*

Also, an analysis using principal axis factoring was conducted in SPSS to examine whether several components can be identified among the criteria in the rubric. The output indicated that there was only one large factor with an eigenvalue of 2.59, accounting for 54% of the variance. The results of the scree plot and parallel analysis also confirmed this finding. No other large components were identified with eigenvalues above 1, indicating that the rubric was unidimensional and that the criteria in the rubric were all working together to measure one underlying trait.

Table 2. *Exploratory Factor Analysis (Principal Axis Factoring)*

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.598	64.946	64.946	2.155	53.868	53.868
2	.688	17.197	82.143			
3	.378	9.451	91.595			
4	.336	8.405	100.000			

Once the reliability of the rubric is confirmed, individual indices as units of analysis should be taken into account to analyze data on the halo effect. To determine the halo effect via individual-level statistics, the rater MS fit indices should be much greater than 1.0 or

much smaller than 1.0, depending on the similarities/dissimilarities of criteria difficulty (Myford & Wolfe, 2004).

To determine the halo effect, three indicators should be considered: the fixed chi-square, trait separation index, and reliability of the trait separation index. The fixed chi-square tests the hypothesis of whether all traits share the same degree of difficulty measure or not. A non-significant value may indicate a halo effect in the ratings of all raters. The trait separation index indicates the number of measurably different levels, or strata, of trait difficulty, and a lower trait separation index implies the halo effect in the ratings. The reliability of the trait separation index provides information about how well a rater can distinguish the items in terms of their levels of difficulty. The ideal of this index is 1 and a low trait separation may connote a halo effect.

Table 3. *Rating scale Measurement Report*

Criteria	Measurement Logit	Model Error	Infit		Outfit	
			MnSq	ZStd	MnSq	ZStd
Grammatical Range and Accuracy	.37	.16	1.01	.1	1.00	.0
Lexical Resource	.03	.16	.79	-1.8	.76	-2.0
Cohesion and Coherence	-.13	.16	.89	-.9	.87	-1.0
Task Achievement	-.26	.16	1.26	2.1	1.29	2.1
M	.00	.16	.99	-.2	.98	-.2
SD	.27	.16	.20	1.7	.23	1.7

RMSE: .16; Adj (True) S.D.: .21; Separation: 1.35; Reliability: .65; Fixed chi-square: 8.5 (*d.f.* = *3; p* = .04)

Table 3 indicates criterion difficulty estimates (measurement logits) along with fit statistics for the four criteria on the writing test rubric. The logit values and fit statistics show the relative difficulty of the criteria and the degree to which all four criteria tap onto the same latent variable. Higher logit values show greater criteria difficulty, while lower logit values indicate lesser criteria difficulty. The criteria are ranked from the most to least difficult in Table 3.

The most difficult criterion was grammatical range and accuracy (.37 logit) followed by lexical resource (.03 logit) and cohesion and coherence (-.13 logit), and the easiest criterion

was task achievement (-.26 logit). In other words, the test-takers had the greatest challenge in achieving a high score on grammatical range and accuracy, while experiencing the least difficulty in getting a high score on task achievement. However, to check if the differences in the difficulty estimates are meaningful, three indices of separation, reliability, and chi-square statistics need to be considered. The separation index indicates the number of statistically distinct levels of criterion difficulty without considering extreme outliers (Linacre, 2018). The possible range of the reliability index is from 0 to 1. When the criteria have similar levels of difficulty, the index is close to 0. In contrast, if the criteria have different levels of difficulty, the index approaches 1. Therefore, the separation index of 1.35 with a reliability of .65 indicates that the four criteria had almost two different levels of difficulty. This finding was supported by the chi-square statistics, which tested a null hypothesis that the difficulty of the four criteria was the same. The chi-square statistic ($\chi 2 = 8.5$, df = 3, p = .01) rejected the null hypothesis, meaning that at least two criteria were significantly different in their difficulty. As aforementioned, a non-significant chi-square means a halo effect, but in the present study, the chi-square is statistically significant. Since the criterion difficulties relatively varied, it was supposed that the MRFM expected ratings would show greater variability (Myford & Wolfe, 2004). Consequently, the scores assigned by the raters who indicated halo effects would not be similar to the expected scores. In other words, the raters' infit and outfit mean-square indices would be much greater than 1 (Myford & Wolfe, 2004). However, as can be observed in Table 4, infit and outfit mean square indices for raters except rater 3 were close to 1, showing that halo effects were not observed in any of these raters' ratings. But in the case of rater 3, infit and outfit mean square indices are more than 1.

Table 4. *Rater Measurement Report*

Criteria	Measurement Logit	Model Error	Infit		Outfit	
			MnSq	ZStd	MnSq	ZStd
1	1.52	.18	.84	-1.3	.82	-1.3
2	.35	.18	1.56	3.6	1.56	3.3
3	-.09	.18	.62	-3.3	.59	-3.3
4	-1.38	.18	.86	-1.1	.87	-1.0
5	-2.1	.18	1.07	.5	1.06	.4
M	-.32	.18	.99	-.3	.98	-.4
SD	1.40	.18	.36	2.6	.36	2.5

RMSE: .18; Adj (True) S.D.: 1.39; Separation: 6.21; Reliability: .98; Fixed chi-square: 3.9 (*d.f. = 3; p* = .00)

Overall, the results showed that the raters did not exhibit halo effects in their ratings, considering the results about the rater fit statistics relative to the criterion difficulty of the rubric.

Discussion and Conclusion

The main focus of the present study was detecting the halo effect across criteria difficulty. To the best of researchers' knowledge, this was the first study conducted to investigate the halo effect across criteria difficulty in a performance-based assessment of writing. The results of a four-facet analysis suggested that except for one rater, no sizeable halo effect across all four criteria was detected based on the acceptable fit statistics and halo indices.

The halo effect is defined as the carry-over from one judgment to another or assigning similar ratings to test takers across items. That raters in this study did not show the halo effect reflects the very fact that raters can distinguish between conceptually distinct criteria of the rubric, which connotes that the rubric generally was functioning properly with raters. This finding is reasonable because the raters rated the test essays criterion-by-criterion (Shin & Ewert, 2015). On the whole, the findings of the present study are in line with those of previous studies (Engelhard Jr, 1994; Kozaki, 2004; Knoch et al., 2007; Yorozuya & Oller Jr, 1980). Engelhard Jr (1994) found that two out of his 15 highly trained raters indicated the halo effect although he did not explain why such a halo effect occurred with two highly trained raters. In an attempt to compare generalizability theory with many-facet Rasch measurement in determining the halo effect, Kozaki (2004) also found that two out of his four professional raters showed signs of the halo effect on two criteria of grammar and vocabulary. She concluded that this halo effect occurred because of the powerful roles these two criteria played in the assessment and judges' overall impression of competence.

Another study conducted by Knoch et al. (2007) found that the halo effect existed due to a lack of training and feedback from her raters. After receiving training and feedback, at least some of the raters did not indicate the halo effect in the face-to-face group although the halo effect remained with the online group raters even after training and feedback.

The results of the present study could be justified for the following reasons with an eye to the previous studies in the field. First, it should be noted that all the raters recruited for the purpose of the present study are considered experienced raters, having many years of experience. Consequently, there is no surprise that they did not show the halo effect except for one rater. Second, each rater had participated in a specialized training program in IELTS

centers in which the writing rubric and the process of rating were elaborated. Training could be of great help for the raters to improve their rating and show fewer signs of the halo effect which is totally in line with the study by Knoch et al. (2007). But in the case of the rating behavior of rater 3, it could be justified in the fact that even experienced raters may show bias toward the rubric criteria. Although rater training has been efficient in reducing rater variance, it is not necessarily effective for all the raters to the same degree. Trained raters sometimes indicate unexpectedly harsh or lenient ratings to rubric criteria. While training may help all raters to become better raters, it cannot eliminate rater subjectivity (Eckes, 2008, 2012; Weigle, 1998). Consequently, any high-stakes decision based on human raters' judgment on performance-based assessment of writing should take into account the effect of rater variance, especially the halo effect.

The halo effect as a subcomponent of raters' variability could be a source of construct irrelevant variance in test scores which adversely affects examinees' scores. It is a rater error in which raters fail to distinguish distinct criteria across a rubric. So, it has to be minimized as much as possible through detailed training and experience on rating. Though, it should be mentioned that this potential error could be reduced rather than eliminated. Pedagogically, through a detailed facet analysis of interactions between raters and rubric criteria, it becomes feasible to detect sources of raters' variance including the halo effect. Based on this information, rater trainers may better realize how to design effective training programs to minimize raters' variance and improve the overall objectivity of human rating (Lumley & Mcnamara, 1995; Schaefer, 2008).

Generalizing the findings of the present study to other contexts should be done with caution. The findings of the present study are restricted to only five trained raters in writing assessments. Other studies should be conducted with more trained raters to justify whether similar findings will be obtained. Furthermore, since only an argumentative type of essay was explored, more research is needed to illustrate whether raters will show the halo effect toward other types of essays as well.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory.* Long Grove, Illinois: Waveland Press.

Anderson, J. R. (2015). *Cognitive psychology and its implications*. Eighth Edition. San Francisco: Worth Publishers.

Andrich, D., Humphry, S. M., & Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the Rasch model. *Journal of Applied Psychological Measurement*, *36*(4), 309–324.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Journal of Applied Psychological Measurement*, *34*(3), 181-196.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. New York: Springer.

Ary, D., Jacobs, L. C., Sorensen, C. K., & Walker, D. (2014). *Introduction to research in education*. Ninth Edition. Wadsworth: Cengage Learning.

Ballard, L. (2017). *The effects of primacy on rater cognition: An eye-tracking study.* Unpublished doctoral dissertation. Michigan State University.

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp.1-22). Wiley Blackwell.

Bechger, T. M., Maris, G., & Hsiao, Y. (2010). Detecting halo effects in performance-based examinations. *Journal of Applied Psychological Measurement*, *34*(8), 607–619.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Brown, A. (1995). The effect of rater variables in the development of an occupation specific language performance test. *Language Testing*, *12*(1), 1-15.

Brown, H. D. (2003). *Language assessment: Principles and classroom practice*. London: Longman.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Second Edition. Boston, Massachusetts: Heinle & Heinle.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*(1), 67-96.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185.

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43-73). Lang.

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

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Second Edition. Bern, Switzerland: Peter Lang.

Engelhard Jr, 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.

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust, & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques,* (pp. 83–102). Routledge.

Hamp-Lyons, L., & Davies, A. (2008). The Englishes of English test: Bias revisited. *World English*, *27*(1), 26-39.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge University Press.

Jeong, H. (2017). Narrative and expository genre effects on students, raters, and performance criteria. *Assessing Writing*, *31,* 113-125.

Kim, S. C., & Wilson, M. (2009). A Comparative Analysis of the Ratings in Performance Assessment Using. *Journal of Applied Measurement*, *10*(4), 403-423.

Knoch, U., Read, J., & Von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training?. *Assessing Writing*, *12*(1), 26–43.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*(1), 1–27.

Lai, E. R., Wolfe, E. W., & Vickers, D. (2015). Differentiation of illusory and true halo in writing scores. *Journal of Educational and Psychological Measurement, 75*(1), 102–125.

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*, 332-340.

Li, J., & Huang, J. (2022). The impact of essay organization and overall quality on the holistic scoring of EFL writing: Perspectives from classroom English teachers and national writing raters. *Assessing Writing*, *51*, 136-150.

Linacre, J. M. (2010). *A user's guide to facets: Rasch-model computer programs*. Facets Rasch measurement computer program [software manual], version 3.67.0. Winsteps.com.

Linacre, J. M. (2016). *Facets Rasch measurement computer program*. Chicago: Winsteps.com.

Linacre, J. M. (2018). *A user's guide to FACETS: Rasch-model computer program*. Chicago: Winsteps.com.

Lumley, T. (2006). *Assessing second language testing: The rater's perspective*. Bern, Switzerland: Peter Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180.

Marais, I., & Andrich, D. (2011). Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, *12*(3), 194-211.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

McNamara, T. F. (1996). *Measuring second language performance*. Harlow, United Kingdom: Longman Publishing Group.

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

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

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

Ockey, G. J. (2012). Item response theory. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 316-328). Routledge.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor & P. Falvey (eds.*), IELTS collected papers. Research in speaking and writing performance* (pp. 446-478). Cambridge University Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: Chicago University Press.

Schaefer, R.T. (2008) *Racial and Ethnic Groups.* Eleventh Edition, Pearson Education, 69.

Shaw, S., & Weir, C. J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing, Studies in Language Testing 26*. UCLES/Cambridge University Press.

Shin, S. Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing*, *32*(2), 259-281.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239-261.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25-29.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind. In L. H. Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Ablex.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.

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

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* London: Palgrave Macmillan.

Yorozuya, R., & Oller Jr, J. W. (1980). Oral proficiency scales: Construct validity and the halo effect. *Language Learning*, *30*(1), 135-153.

Zhang, X., & Lu, X. (2022). Revisiting the predictive power of traditional vs. fine-grained syntactic complexity indices for L2 writing quality: The case of two genres. *Assessing Writing*, *51*, 106-122.