Language Testing in Asia

RESEARCH Open Access

Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis

Hyunwoo Kim

Correspondence: kim01@snu.ac.kr
Department of English Language Education, College of Education, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea

Abstract

The halo effect is raters' undesirable tendency to assign more similar ratings across rating criteria than they should. The impacts of the halo effect on ratings have been studied in rater-mediated L2 writing assessment. Little is known, however, about the extent to which rating criteria order in analytic rating scales is associated with the magnitude of the group- and individual-level halo effects. Thus, this study attempts to examine the extent to which the magnitude of the halo effect is associated with rating criteria order in analytic rating scales. To select essays untainted by the effects of rating criteria order, a balanced Latin square design was implemented along with the employment of four expert raters. Next, 11 trained novice Korean raters rated the 30 screened essays with respect to the four rating criteria in three different rating orders: standard-, reverse-, and random-order. A three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty) was fitted to estimate the group- and individual-level halo effects. The overall results of this study showed that the similar magnitude of the group-level halo effect was detected in the standard- and reverse-order rating rubrics while the random presentation of rating criteria decreased the group-level halo effect. A theoretical implication of the study is the necessity of considering rating criteria order as a source of construct-irrelevant easiness or difficulty when developing analytic rating scales.

Keywords: Halo effect, Rating criteria order, Rating scale model (RSM), Many-facet Rasch measurement (MFRM)

Introduction

Background of study

The halo effect is defined as rater's cognitive bias, where the judgment of a certain rating criterion is influenced by that of related other rating criteria of test takers' performance. A body of research has examined the halo effect in the context of rater-mediated performance assessment (Andrich, Humphry, & Marais, 2012; Andrich & Kreiner, 2010; Bechger, Maris, & Hsiao, 2010; Engelhard, 1994; Lai, Wolfe, & Vickers, 2015; Marais & Andrich, 2011; Myford & Wolfe, 2003, 2004). More specifically, in operational scoring sessions, the halo effect induced by muted raters manifests itself as the unduly flattened profile of ratings, thereby significantly limiting the value of diagnostic feedback given to test takers and nullifying the additional efforts in developing multiple rating criteria. The sources of the halo effect have been known to entail rater's general impression, a salient rating criterion, and an inability of raters mainly induced by insufficient rater training (Lance, Lapointe, & Fisicaro, 1994).

However, aside from these three sources of the halo effect, design features of analytic rating scales as the source of the halo effect have been relegated to a lesser position, although rating criteria order as an underlying mechanism of the halo effect has been suggested in rater-mediated performance assessment (Balzer & Sulsky, 1992; Fisicaro & Lance, 1990; Judd, Drake, Downing, & Krosnick, 1991; Murphy, Jako, & Anhalt, 1993). More specifically, the impact of design features on the halo effect was empirically demonstrated (Humphry & Heldsinger, 2014), and the results of eye-movement research on rater cognition showed that rating criteria order was associated with the rating process and rater severity in L2 writing assessment (Ballard, 2017; Winke & Lim, 2015). Regarding the effects of rating criteria order on the halo effect, Lai et al. (2015) clearly stated the need to control the sequence in which rating criteria are rated to identify which rating criteria are most vulnerable to the halo effect in L2 writing assessment.

To address the call for continued research on the halo effect, the aim of this study is to investigate the extent to which the magnitude of the halo effect is associated with the manipulation of rating criteria order as a design feature of analytic rating scales. More specifically, this study attempts to examine the extent to which the invariance of rating criterion difficulty is compromised due to the halo effect by using many-facet Rasch measurement. To achieve this aim, when trained novice Korean raters rated the essays untainted by the effects of rating criteria order, the magnitude of the halo effect was investigated with respect to three different rating orders: standard-, reverse-, and random-order. In the standard-order analytic rating rubric, rating criteria were successively displayed to the raters as follows: *content*, *organization*, *vocabulary*, and *language use*. On the other hand, this order was precisely reversed in the reverse-order analytic rating rubric. Rating criteria were randomly presented to the raters one after the other in the random-order analytic rating rubric. A three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty) was separately fitted to estimate the group- and individual-level halo effects.

Literature review

Validity of ratings

In their seminal paper, Cronbach and Meehl (1955) presented three distinct types of validity: criterion-oriented validity, content validity, and construct validity. Criterion-oriented validity is concerned with the relationship between a test and a criterion in which predictions about test takers are made and can be further broken down into predictive and concurrent validity. The differences between the two types of validity lie in whether the prediction is made for the future or present criteria. Content validity can be enhanced when the content of a test is shown to be a representative sample of the domain of a test. Lastly, construct validity can be defined as the degree to which test scores reflect the targeted ability.

Criticizing Cronbach and Meehl's three types of validities (1955), Messick (1989) suggested that criterion-oriented and content validities should be subsumed under construct validity. He thus reconceptualized validity as a unitary entity and defined it as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Drawing on Messick's conceptualization, in this study, the validity of ratings is defined as the extent to which inferences about L2 writing ability are justified based upon ratings.

In rater-mediated L2 writing assessment, the sources of threats to the validity of ratings entail both rater variability and rating scales. Firstly, rater variability is defined as "systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee [test taker]" (Scullen, Mount, & Goff, 2000, p. 200). Secondly, rating scales themselves could pose a threat to the validity of ratings when a poorly constructed rating rubric fails to adequately reflect the target ability (Messick, 1994). For example, the omission of important rating criteria to measure L2 writing ability leads to construct under-representation as all aspects of the construct are not tapped by rating scales.

More importantly, the inconsistent interaction between raters and rating scales could introduce systematic construct-irrelevant variances when raters interpret rating criteria in a way a rating criterion is not intended and become influenced by irrelevant aspects of the essays, such as the neatness of test takers' handwriting. In this regard, the systematic bias between raters and rating scales could introduce construct-irrelevant easiness or difficulty from the test takers' perspective by disrupting the hierarchy of rating criterion difficulty (i.e., the invariance of rating criterion difficulty), which pertains to a specific form of validity evidence (AERA, APA, & NCME, 2014).

Rater bias and rater training

In rater-mediated L2 writing assessment, the inconsistent interaction between raters and rating scales manifests itself as empirically distinct rater biases, including differential severity or leniency, centrality or restriction of range, and the halo effect (Eckes, 2015; Engelhard, 1994; Myford & Wolfe, 2003, 2004). Rater severity or leniency is a consistent tendency of raters to award higher or lower ratings to test takers' writing performance than the test takers deserve. The central tendency of raters indicates the consistent tendency for raters to intentionally avoid awarding extreme scores of rating criteria in analytic rating scales. Although the range restriction of ratings is similar to centrality, the range restriction of ratings tends to materialize when ratings cluster around any point in the analytic rating scales (Saal, Downey, & Lahey, 1980). The halo effect is conceptualized as "putatively inflated correlations" (Cooper, 1981, p. 219) among subconstructs of L2 writing ability.

In L2 writing assessment, the logical course of action is to control and minimize, if not prevent, rater inconsistency, a major threat to the validity of ratings. Bachman and Palmer (1996) suggest that a major source of rater biases is insufficient rater training, and there has been a wide consensus in the L2 writing assessment literature on the effectiveness of rater training at reducing rater biases among newly recruited novice raters (Weigle, 1994, 1998).

Although rater training has been efficient in reducing rater biases among inexperienced raters, it is harder to get experienced raters to reformulate their beliefs about effective writing regardless of the types of analytic rating scales they use to rate the essays. In other words, once raters internalize the specific rating rubric and formulate their own beliefs about effective writing, they may tend to harbor unwavering notions about the relative importance of rating criteria, leading to systematic rater bias. Consistent with this suggestion, Weigle (1994) showed that judgments made by experienced raters were less affected by rater training than judgments made by inexperienced raters. Additionally, it has been empirically demonstrated how experienced raters' belief about the relative importance of each rating criterion in analytic rating scales led to heightened severity or leniency towards a specific rating criterion (Eckes, 2008, 2012).

Halo effect

The halo effect is defined as rater's cognitive bias where the judgment of a certain rating criterion is influenced by that of related other rating criteria of test takers' performance. The halo effect manifests itself as the erroneously inflated intercorrelations among distinct rating criteria in analytic rating scales (Myford & Wolfe, 2003). When raters are subject to the halo effect, they tend to award more similar or even identical ratings across rating criteria than they should.

When investigating literary qualities of imaginative American writer, Wells (1907) initially proposed the halo effect in which the distinct literary traits of the authors "tended to be interpreted rather regarding general merit" (p. 21). Since this first proposal, Fisicaro and Lance (1990) have suggested three causal models for the halo effect to occur: the general impression model, the salient dimension model, and the inadequate discrimination model. According to the general impression model, the halo effect is regarded as the influence of the overall impression of the performance upon raters' judgment of other aspects of the performance. More specifically, the general impression model explains the sources of the halo effect when raters are asked to assign holistic ratings to the essay before assigning analytic ratings to rating criteria in analytic rating scales. For example, according to this model, the halo effect may be exacerbated when overall ratings such as *overall task fulfillment* coexist in analytic rating scales with other rating criteria such as grammar when rating test takers' writing ability (Mc-Namara, 1990).

The salient dimension model suggests that raters' judgment of more salient rating criteria of the essays could affect the raters' subsequent judgment of less salient aspects of the essays. For example, in the ESL composition profile developed by Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey (1981), the differential weightings are assigned to *content* (30), *organization* (20), *vocabulary* (20), *language use* (25), and *mechanics* (5) so that the total of differential weightings across rating criteria equals 100. In this situation, it is obvious that raters consider the rating criterion with a greater weight to be more important than others.

In the inadequate discrimination model, the halo effect occurs when raters are incapable of discriminating between distinct dimensions of the performance or when the scoring descriptors for rating criteria are unclear (Cooper, 1981; Knoch, 2009; North & Schneider, 1998; Upshur & Turner, 1995). More specifically, when raters are not able to differentiate between rating criteria either within a specific test taker or regardless of test takers, the halo effect would occur. Lexical complexity or sophistication of the essays could influence ratings upon two rating criteria, such as *vocabulary* and *language use*, thus inflating the correlation between two rating criteria when raters are incapable of differentiating between the syntactic and lexical complexity of the essays. The unclear scoring descriptors in rating criteria in analytic rating scales also contribute to the magnitude of the halo effect.

Although the halo effect has been regarded as one of the serious rater errors (Engelhard, 1994; Fisicaro, 1988; Myford & Wolfe, 2003), it is worthwhile to note that the influences of the halo effect on the accuracy of ratings seem equivocal. Even though the greater the halo effect is, the lower the accuracy of ratings becomes, the halo effect could paradoxically enhance the accuracy of the ratings (Fisicaro, 1988). For example, through a meta-analysis of studies on rater errors, Murphy and Balzer (1989) disconfirmed the hypothesis that the size of the halo effect is negatively associated with some accuracy measures in ratings. In their study, studies on rater errors, including the halo effect, leniency, and range restriction of ratings, were analyzed to examine the population correlation between the six measures of three rater errors (halo, leniency, and range restriction) and four accuracy measures. After correcting sampling and measurement errors, the halo effect was found to be weakly associated with the accuracy measures of ratings, a conundrum that is referred to as the “halo-accuracy paradox” (Cooper, 1981, p. 240). In this regard, Murphy et al. (1993) rightly pointed out that the halo effect does not always decrease the accuracy of ratings depending on the purposes of ratings. For example, when the primary purpose of ratings is to rank order test takers based on the overall performance, the halo effect could potentially lead to the enhanced reliability of ratings due to the increased intercorrelations among rating criteria.

Although the influences of the halo effect on the accuracy of ratings seem equivocal (Fisicaro, 1988; Murphy & Balzer, 1989), the erroneously flattened profile of ratings prompted by the halo effect (e.g., all the same ratings such as 1 s or 4 s across rating criteria) is problematic in that “ESL writers often acquire different components of written control at different rates” (Hamp-Lyons, 1991, p. 241). In other words, the erroneously uniform ratings provoked by rater’s cognitive bias presumably make it difficult to interpret test takers’ strengths and weaknesses regarding subconstructs of L2 writing ability and to measure the developmental trajectory across subconstructs of L2 writing ability. Consequently, the halo effect limits the value of feedback L2 writers receive from their analytic scores.

Many-facet Rasch measurement of halo effect

There is a growing consensus in the Rasch measurement literature about how to detect the presence of the halo effect (Eckes, 2012, 2015; Engelhard, 1994; Myford & Wolfe, 2004), and a few empirical studies have investigated the magnitude of the halo effect using many-facet Rasch measurement (Farrokhi & Esfandiari, 2011; Knoch, Read, & von Randow, 2007; Kozaki, 2004; Schaefer, 2008). More specifically, when the sizable halo effect persists, rating criteria are overfit to the model expectations as a score awarded to a rating criterion is too predictable to determine from scores awarded to other rating criteria. In this situation, fit statistics of rating criteria are consulted to estimate the magnitude of the halo effect.

When the less serious halo persists, the difficulty of rating criteria would be rendered similar to each other. In this situation, group-level indicators, such as the fixed chi-square statistic, separation index, and reliability of separation, assist in the detection of the halo effect exhibited by raters as a group. For example, the fixed chi-square statistic is used to test the hypothesis that all rating criteria in analytic rating scales share the same level of difficulty measures after accounting for measurement error. In this situation, the non-significant chi-square statistic suggests that raters are not able to differentiate between conceptually distinct rating criteria (Myford & Wolfe, 2004). Similarly, the separation index indicates the statistically distinct levels of difficulty after accounting for measurement error, and the reliability of separation represents the extent to which raters reliably differentiate between the statistically distinct levels of difficulty. According to Linacre (2018a), only when the separation index is greater than 3.00 and the reliability of separation is greater than .90, the hierarchy of rating criteria is established, and thus the validity of a test instrument with respect to the internal structure of rating criteria is supported.

In order to detect the halo-exhibiting raters, rater fit statistics are consulted to determine muted raters who are subject to the halo effect (Eckes, 2015; Engelhard, 1994; Myford & Wolfe, 2004). As the difficulty measures of rating criteria might vary, the interpretation of rater fit statistics is quite context-bound as it largely depends on the variation of rating criterion difficulty measures. To illustrate when the difficulty estimates of rating criteria vary slightly, muted raters whose fit statistics are significantly less than 1.00 are considered to exhibit the halo effect in that the expected ratings by the model are similar across rating criteria. On the other hand, when the difficulty estimates of rating criteria vary greatly, the expected ratings by the model across rating criteria vary as well. In this situation, misfitting raters whose fit statistics are significantly more than 1.00 are muted raters exhibiting the halo effect, as the expected ratings by the model across rating criteria display considerable variation.

Aside from rater fit statistics, there have been also attempts to operationalize the halo effect as the magnitude of local dependence among rating criteria exhibited by individual raters (Andrich & Marais, 2019), and this approach is particularly useful when all the raters exhibit a common sizable halo effect (Marais & Andrich, 2011). More specifically, the magnitude of local dependence is operationalized as the degree of the increased or decreased difficulty measures of the dependent dichotomous items (Andrich

& Kreiner, 2010) or the degree of the shift of thresholds of the dependent polytomous items (Andrich et al., 2012). Although this approach sounds promising, it is critical to note that the significance of the local dependence between the independent criterion and its dependent criterion requires a large sample size that is often unavailable in language assessment studies.

Rating criteria order

Rating criteria order has been suggested as an underlying mechanism of the halo effect in rater-mediated performance assessment (Balzer & Sulsky, 1992; Judd et al., 1991; Murphy et al., 1993). More specifically, it is suggested that the judgments or impressions about one rating criterion could spread similar impressions to other rating criteria, thus potentially resulting in similar ratings across rating criteria. In the context of rater-mediated L2 writing assessment, two important eye-movement studies of raters have shed more light on how the order of rating criteria in analytic rating scales is associated with rater perception (Ballard, 2017; Winke & Lim, 2015).

To illustrate, Winke and Lim (2015) empirically demonstrated that the left-most rating criterion in analytic rating scales, *content* in their study, elicited the largest amount of attention from raters, as evidenced by the eye-fixation duration. On the other hand, the right-most subconstruct, *mechanics*, elicited the least amount of attention from raters. Similarly, Ballard (2017) demonstrated that rating criteria order in analytic rating scales elicited the differential attention to rating criteria, which, in turn, was found to affect the magnitude of interrater reliability and severity of raters.

Although these two studies did not fully address how the order of rating criteria in analytic rating scales is associated with the magnitude of the halo effect, rating criteria order was found to affect the amount of attention that raters pay to each rating criterion. In this respect, Ballard (2017) and Winke and Lim (2015) indirectly suggest that the halo effect might be associated with rating criteria order in analytic rating scales in that a certain rating criterion could be artificially rendered salient among other rating criteria by presenting a specific rating criterion first to raters.

Purpose

Although the halo effect has been shown to be a major source of rater errors in L2 writing assessment, it is apparent that the magnitude of the halo effect induced by rating criteria order has been underexplored. In order to occupy this research gap, the current study attempts to examine the magnitude of the halo effect exhibited by trained novice Korean raters by implementing a three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty).

In this study, the study population is trained novice Korean raters. This is mainly because novice raters are more subject to the salience of rating criteria induced by the variation of the order of rating criteria in analytic rating scales than experienced raters in that novice raters have not yet formulated their idiosyncratic beliefs about the relative importance of rating criteria.

In order to estimate the extent to which rating criteria order is associated with the magnitude of the halo effect exhibited by trained novice raters, it is necessary for the trained novice Korean raters in the study to rate the essays that appropriately display uneven profiles across four rating criteria (i.e., essays untainted by the effects of rating criteria order). Otherwise, the resultant profiles of the essays are confounded with both the effects of rating criteria order and halo induced by novice Korean raters' cognitive biases. To achieve this aim, 33 essays from the International Corpus Network of Asian Learners of English (ICNALE) were initially screened by the researcher in the study (Ishikawa, 2013). In turn, to further dissipate the effects of rating criteria order, the 33 essays were rated by four expert raters with the implementation of a balanced Latin square design.

Based on the results of the previous eye-movement studies (Ballard, 2017; Winke & Lim, 2015), three types of analytic rating scales were created in this study: standard-, reverse- and random-order analytic rating scales. In the standard-order analytic rating rubric, rating criteria were presented to the raters in the following order: *content*, *organization*, *vocabulary*, and *language use*. In the reverse-order analytic rating rubric, this order was precisely reversed. In the random-order analytic rating rubric, rating criteria were randomly presented to the raters. Additionally, to control the direction of the influence of the immediately preceding rating criterion, raters were unable to revisit the previous rating criterion to revise scores. The independent variable in the study is three presentation sequences of rating criteria orders in analytic rating scales. The dependent variable includes Rasch-based halo indices available in a three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty).

The main objective of this study is to answer the following research question:

- To what extent is rating criteria order in analytic rating scales associated with the magnitude of the group- and individual-level halo effects exhibited by trained novice Korean raters?

Method

Expert and novice raters

In order to select essays untainted by the effects of rating criteria order, four experienced ESL instructors at a large state university in the USA were recruited. The instructors' experience in teaching ESL writing courses totaled a minimum of 3 years at the time when this study was conducted, and they were quite familiar with Jacobs' analytic rating rubric (Jacobs et al., 1981), as its modified analytic rating rubric had been implemented in the English placement test at the same state university.

To estimate the magnitude of the halo effect induced by rating criteria order, the researcher recruited 12 novice Korean raters with three males and nine females, aged between 26 and 45; they were master's students or in-service teachers in secondary school or private educational institutions in South Korea whose major was associated with teaching English as a foreign language (TEFL) at the time when the study was conducted. Only those who had never rated essays with any variants of the analytic rating rubric originally developed by Jacobs et al. (1981) were eligible to participate in the study.

Essays

The researcher in this study selected 33 argumentative essays from the International Corpus Network of Asian Learners of English (ICNALE), the topic of which is about *smoking in restaurants* (Ishikawa, 2013). It should be important to note that those essays were originally collected for the purpose of building a learner corpus. As the original ratings of those essays were available in the corpus, the researcher initially selected only those 33 essays which did not display the flat profiles across rating criteria.

Revised analytic rating scales

The widely implemented analytic rating rubric in the context of ESL writing assessment was modified to answer the research questions in the study (Jacobs et al., 1981). Firstly, it was decided that the rating criterion of mechanics would be excluded, as few scoring descriptors in the original category (e.g., illegible handwriting) are not relevant to the study; all the essays had been transcribed and digitally stored as text files prior to analysis. However, some of the descriptors of mechanics, such as spelling errors, were incorporated into the vocabulary rating criterion, as they were deemed relevant aspects of L2 writing ability (see Appendix Table 14). Furthermore, punctuation errors that might influence the judgment of other aspects of the essays had been corrected by the researcher in the study prior to the data collection. Secondly, nominal weightings to each rating criterion were deleted in order not to distort the perceived importance of rating criteria, and 25- or 30-point scales of rating criteria collapsed into a 7-point scale to enhance the reliability of ratings given by raters.

Procedure

Expert online rating session

Four expert raters rated the 33 essays online using the revised analytic rating scales. To obtain ratings untainted by the effects of rating criteria order, only a single rating criterion was rated at a time. More importantly, the order of rating criteria was counterbalanced with the implementation of a balanced Latin square design (Maxwell & Delaney, 2004), as shown in Table 1. Additionally, to prevent raters from remembering their previous ratings on the same essay, the order of essays was shuffled at each online rating session, and at least a 1-week-long interval was maintained between online rating sessions. Three out of the 33 essays were used as benchmark essays in the rater training session for the recruited Korean raters. The remaining 30 essays were used as the essays for three online rating sessions where the recruited Korean raters were asked to use three types of analytic rating scales.

Table 1 Rating criteria order for expert raters

Rater	Rating criteria			
01	Content	Organization	Language use	Vocabulary
02	Organization	Vocabulary	Content	Language use
03	Vocabulary	Language use	Organization	Content
04	Language use	Content	Vocabulary	Organization

Novice online rating session

Before the data collection, 12 novice Korean raters participated in both face-to-face and online rater training sessions. During the face-to-face training session, the scoring descriptors for each rating criterion were clarified with examples provided to illustrate each. During the online rater session, the trained novice Korean raters rated three benchmark essays online that were derived from the previous expert rating sessions. Ratings provided by the novice raters were compared with those fair-average scores computed from ratings of the four expert raters. When huge discrepancies between the ratings arose, further rater training was conducted until recruited Korean raters were deemed to qualify as a reliable rater.

Three types of analytic rating scales were developed using *Qualtrics*, the web-based survey tool for the trained novice Korean raters: the standard-, reverse-, and random-order analytic rating scales. The *Qualtrics* user interface was designed in a way that raters had to evaluate each essay in a predetermined order and were unable to correct their scores for a previous rating criterion. To illustrate, all rating criteria were not simultaneously displayed to raters, but only one rating criterion at a time was presented to raters along with the essay under evaluation.

Next, 12 raters were randomly assigned to three rating groups, as shown in Table 2, and then participated in three online rating sessions accordingly, where they used different analytic rating scales. To prevent raters from remembering their previous scores on the same essay, the presentation order of the 30 essays was shuffled for every rating session. Additionally, raters were not allowed to move on to the next online rating sessions before less than 1 week from the completion of the previous rating session.

Results

Expert ratings

The interrater reliability of the expert raters was checked using an intraclass correlation coefficient (ICC), as shown in Table 3. As shown in the off-diagonal statistics, weak to moderate intercorrelations existed between rating criteria. Interrater reliabilities across the four rating criteria were satisfactory, as shown in diagonal statistics in the correlation matrix.

Group-level halo effect

A three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty) was fitted to estimate the magnitude of the group-level halo effect exhibited by the four expert raters using FACETS (Ver 3.82). The major assumptions of the rating scale model were checked, following the suggestions by Fan and Bond (2019). As illustrated in the acceptable rating criterion fit statistics and point-measure biserial correlations of the four rating criteria in Table 4, the assumptions of unidimensionality

Table 2 Rating scale order for novice raters

Group		Rating scale	
01	Standard	Reverse	Random
02	Reverse	Random	Standard
03	Random	Standard	Reverse

Table 3 Correlations for expert ratings

	M	SD	1	2	3	4
1. Content	3.53	1.75	.82			
2. Organization	2.83	1.75	.61**	.84		
3. Vocabulary	4.36	1.56	.37**	.23**	.82	
4. Language use	4.50	1.63	.29**	.35**	.51**	.87

Note. Interrater reliability is an intraclass correlation coefficient (ICC [3, 4])
***p* < .01, two-tailed

and local independence seemed reasonably defensible. The results of the principal component analysis of Rasch standardized residuals further supported the assumption of unidimensionality; the eigenvalue of the first principal component of Rasch standardized residuals was equal to 1.57, thus not exceeding the strength of the two items (Linacre, 2018b).

Table 4 also shows indices of the group-level halo effect exhibited by the four expert raters. As evidenced in the heterogenous difficulty measures among the four rating criteria regarding the group-level halo effect, the expert raters as a group did not appear to award more similar ratings across the four rating criteria than they should, as shown in the significant fixed-effect chi-square statistic, $\chi^2(3) = 133.5$, $p < .001$. Additionally, both the separation ratio and index jointly suggested that more than nine statistically distinct levels of rating criterion difficulty among the four rating criteria existed. The four expert raters reliably differentiated the four rating criteria in terms of difficulty when rating the 30 essays, as shown in the reliability of separation, .98.

Individual-level halo effect

As indicated in the acceptable rater fit statistics (between 0.80 and 1.20) in Table 5, it appears that the four expert raters awarded neither noisy nor redundant ratings to the 30 essays except for expert rater 02; however, this rater was only very slightly outside of the rather conservative 0.80 and 1.20 range.

In order to further ensure that the 30 essays correctly displayed the uneven profiles across the four rating criteria, the fair-average scores of each rating criterion were

Table 4 Rasch summary statistics of expert ratings

	Difficulty	Infit	Outfit	PTMEA
Rating criterion				
Content	0.38 (0.08)	1.06	1.05	.70
Organization	0.85 (0.08)	1.02	0.97	.70
Vocabulary	− 0.09 (0.08)	1.02	1.07	.57
Language use	− 0.33 (0.08)	0.89	0.90	.66
Halo index				
Separation ratio	6.64			
Separation index (Strata)	9.18			
Reliability of trait separation	.98			
Fixed-effect (all same) chi-square statistic	133.5 (*df* = 3, *p* < .001)			

Note. Standard errors of measurement are presented in parentheses. *PTMEA* = point-measure biserial correlation

Table 5 Rater fit statistics of expert raters

Expert rater	Severity (SE)	Infit	Outfit	Variance[a]
01	0.50 (0.08)	1.07	1.04	2.42
02	− 0.10 (0.08)	1.24	1.25	2.15
03	− 0.06 (0.08)	0.87	0.87	1.93
04	− 0.34 (0.08)	0.82	0.83	2.65

[a]Average within-test taker variances

computed using FACETS (Ver 3.82). As shown in Table 6, none of the essays displayed the unduly flattened profiles of ratings across the four rating criteria except for essays 02, 16, and 29. Rather than excluding those essays, however, they were included in the subsequent analysis because the sample size of essays needs to be substantial to exhaust all the seven categories in the current rating scale; failure to meet this requirement

Table 6 Fair-average scores of the 30 screened essays

Essay	Ability (logit)	Content	Organization	Vocabulary	Language use
01	1.05	6.01	5.28	4.77	5.39
02	2.66	6.76	6.53	6.03	6.94
03	0.26	2.94	3.78	4.77	5.63
04	0.05	2.68	2.21	4.28	6.56
05	0.47	4.23	3.53	6.03	4.66
06	− 0.44	2.94	2.47	4.03	3.41
07	− 0.40	2.42	2.47	3.52	4.66
08	0.01	3.21	2.21	5.78	4.41
09	0.30	4.48	4.78	3.52	4.66
10	− 0.19	2.94	3.27	3.77	4.41
11	− 0.82	3.21	1.47	1.95	3.92
12	− 0.07	3.47	1.96	4.53	5.14
13	− 0.11	4.23	3.53	2.99	4.17
14	− 0.48	2.17	1.71	3.77	4.90
15	− 0.95	1.68	1.71	3.52	2.88
16	0.39	4.73	4.28	4.03	4.90
17	0.05	3.73	2.47	4.77	4.90
18	− 0.60	2.17	1.47	3.52	4.66
19	− 0.77	2.68	1.71	3.26	3.15
20	1.34	4.73	4.53	6.28	6.94
21	0.26	4.48	2.74	5.28	4.66
22	− 0.31	2.17	2.21	4.28	4.90
23	− 0.11	2.94	1.96	5.02	4.90
24	0.30	1.92	3.00	5.53	6.78
25	0.75	4.73	3.53	5.53	6.11
26	− 0.52	3.21	1.71	4.28	3.15
27	− 0.73	2.94	1.71	2.99	3.41
28	0.22	4.48	3.27	3.77	5.39
29	− 1.26	2.42	1.47	2.21	2.07
30	− 0.35	3.98	2.74	3.52	3.15

Table 7 Correlations for novice ratings (standard)

Rating criterion	*M*	SD	1	2	3	4
1. Content	3.34	1.62	.96			
2. Organization	3.47	1.67	.90**	.97		
3. Vocabulary	3.68	1.35	.60**	.72**	.94	
4. Language use	3.67	1.47	.64**	.67**	.86**	.94

Note. Interrater reliability is an intraclass correlation coefficient (ICC [3, 4])
$^{**}p$ < .01, two-tailed

(e.g., arising from unobserved or null categories) could lead to the unstable estimate of rating criterion difficulty (Linacre, 1999).

Novice ratings

Initially, 12 trained Korean raters rated the 30 screened essays; they rated the four rating criteria in different orders as mandated by the analytic rating rubric design. Despite the rater training sessions, one rater (rater 04) continued providing noisy and unreliable ratings, resulting in rater fit statistics that were greater than 2.00 across three analytic rating rubrics. Thus, the subsequent analysis contained only 11 trained novice Korean raters.

The interrater reliability of the remaining 11 trained Korean raters was checked using an intraclass correlation coefficient, as shown in Tables 7, 8, and 9. Interrater reliabilities across the four rating criteria seemed satisfactory, as evidenced in diagonal statistics in each of the correlation matrices. Additionally, Table 10 illustrates traditional indices of the group-level halo effects, including the average of the between-criterion correlations and that of within-test takers variances.

Group-level halo effect

A three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty) was separately fitted to estimate the magnitude of the group- and individual-level halo effects for each rating criteria order. The major assumptions of the rating scale model were checked, following the suggestions by Fan and Bond (2019). As demonstrated by the acceptable ranges of fit statistics and point-measure biserial correlations of the four rating criteria in Table 11, the assumptions of unidimensionality and local independence seemed to reasonably hold. Additionally, the principal component analysis of Rasch standardized

Table 8 Correlations for novice ratings (reverse)

Rating criterion	*M*	SD	1	2	3	4
1. Content	3.44	1.53	.95			
2. Organization	3.49	1.60	.89**	.96		
3. Vocabulary	3.72	1.46	.73**	.73**	.94	
4. Language use	3.69	1.54	.74**	.73**	.86**	.94

Note. Interrater reliability is an intraclass correlation coefficient (ICC [3, 4])
$^{**}p$ < .01, two-tailed

Table 9 Correlations for novice ratings (random)

Rating criterion	*M*	SD	1	2	3	4
1. Content	3.37	1.43	.95			
2. Organization	3.26	1.48	.86**	.96		
3. Vocabulary	3.67	1.37	.70**	.68**	.94	
4. Language use	3.67	1.40	.64**	.67**	.83**	.94

Note. Interrater reliability is an intraclass correlation coefficient (ICC [3, 4])
***p* < .01, two-tailed

residuals further confirmed that the unidimensionality assumption of the model was met for the standard- and random-order datasets, but not for the reverse-order dataset; in the reverse-order dataset, the first principal component of Rasch standardized residuals amounted to 2.19, which exceeded the eigenvalue of 2.00 (i.e., the strength of two rating criteria). Thus, following the suggestion by Bonk and Ockey (2003), the additional principal component analysis of raw ratings was conducted to verify whether the unidimensionality assumption held using SPSS 22.0.

The results of the principal component analysis of raw ratings indicated that the first principal component of raw ratings was dominant enough to constitute the Rasch dimension, which was interpretable as L2 writing ability of test taker. More specifically, the first principal component (the Rasch dimension) accounted for 83.6% of the total variance with eigenvalues in excess of 3.00, while the eigenvalues of the other three components did not exceed 0.50. Additionally, all four rating criteria loaded highly on the Rasch dimension with loadings in excess of 0.90. As jointly evidenced in the acceptable fit statistics, and the results of the principal component analysis of raw ratings, the unidimensionality assumption for the reverse-order dataset was defensible.

The magnitude of the group-level halo effect exhibited by the 11 trained novice Korean raters is also illustrated in Table 11. When it comes to the standard-order rating rubric, the presence of the serious group-level halo effect was not detected, as shown in the fixed chi-square statistic, $\chi^2(3)$ = 34.5, p < .001. The separation index of 4.68 suggested that more than four statistically distinct levels of rating criterion difficulty existed. The reliability of separation, .92, showed that the 11 trained novice Korean raters reliably differentiated between four distinct rating criteria.

Regarding the reverse-order analytic rating rubric, the sizable group-level halo effect did not seem to persist, either, as evidenced in the significant fixed chi-square statistic, $\chi^2(3)$ = 30.3, p < .001. Additionally, the separation index of 4.36 suggested that more than four statistically distinct difficulty levels existed. The four rating criteria were also

Table 10 Conventional halo indices of novice ratings

Rating session	Average intercorrelation	Average within-test taker variance
Standard	.76	0.66
Reverse	.79	0.56
Random	.74	0.59

Table 11 Rasch summary statistics of novice ratings

	Standard (n = 1320)			Reverse (n = 1320)			Random (n = 1320)		
	Infit	**Outfit**	**PTMEA**	**Infit**	**Outfit**	**PTMEA**	**Infit**	**Outfit**	**PTMEA**
Rating criterion									
Content	1.09	1.07	.80	0.98	0.98	.80	1.03	1.02	.77
Organization	0.91	0.90	.84	0.95	0.93	.83	0.87	0.86	.83
Vocabulary	0.90	0.91	.76	0.97	0.97	.78	0.99	0.99	.76
Language use	1.06	1.06	.75	1.07	1.06	.78	1.08	1.07	.76
Halo index									
Separation ratio	3.26			3.02			4.58		
Separation index (Strata)	4.68			4.36			6.44		
Reliability of separation	.91			.90			.95		
Fixed-effect (all same) chi-square statistic	34.5 (df = 3, p < .001)			30.3 (df = 3, p < .001)			66.1 (df = 3, p < .001)		

Note. PTMEA = point-measure biserial correlation

reliably differentiated by the 11 trained novice Korean raters, as evidenced in the satisfactory reliability of separation, .90.

When the four rating criteria were randomly displayed to the raters, the group-level halo effect slightly decreased. More than six rating criterion difficulty levels emerged, as shown in the separation index of 6.44. Furthermore, compared to the standard- and reverse-order analytic rating rubrics, the 11 trained novice Korean raters seemed to make more reliable distinctions among four rating criteria, as shown in the reliability of separation, .95.

It is important to note that the indices for the group-level halo effect should be interpreted in conjunction with the spread of rating criterion difficulty. In fact, despite the halo indices indicating the absence of the sizable group-level halo effect, the spreads of rating criterion difficulty across three analytic rating rubrics were quite narrow. As can be seen in Table 12, the logit spreads were translated to, at most, 0.50 point in the 7-point rating metric across three analytic rating rubrics

More importantly, as demonstrated in the 95% confidence intervals in Table 12, the rating criteria of *content* and *organization* were similarly difficult, and there were no significant differences between the rating criteria of *vocabulary* and *language use* in

Table 12 Rasch summary statistics of rating criterion difficulty

	Standard		Reverse		Random	
	Difficulty	**95% CI**	**Difficulty**	**95% CI**	**Difficulty**	**95% CI**
Rating criterion						
Content	0.78 (.06)	[0.66, 0.90]	0.68 (.06)	[0.56, 0.80]	0.98 (.06)	[0.86, 1.10]
Organization	0.69 (.06)	[0.57, 0.81]	0.65 (.06)	[0.53, 0.77]	1.10 (.06)	[0.98, 1.22]
Vocabulary	0.39 (.06)	[0.27, 0.51]	0.34 (.06)	[0.22, 0.46]	0.54 (.06)	[0.42, 0.66]
Language use	0.39 (.06)	[0.27, 0.51]	0.33 (.06)	[0.21, 0.45]	0.53 (.06)	[0.41, 0.65]
M	0.56		0.50		0.79	
SD	0.19		0.18		0.29	

Note. Standard errors of measurement are presented in parentheses. *CI* confidence interval terms of criterion difficulty. In other words, the 11 trained novice Korean raters as a group awarded similar ratings to the rating criteria of *content* and *organization* across three rating rubrics; similarly, they awarded similar ratings to the rating criteria of *vocabulary* and *language use*.

Individual-level halo effect

In this study, the spreads of rating criterion difficulty across three analytic rating rubrics were quite narrow. Therefore, to facilitate the detection of halo-exhibiting raters in the study, rating criterion difficulty was anchored at 0.00, following the steps proposed by Linacre (2018b). The rationale behind the steps is that individual raters who are overfit to the model expectations (< 0.80) are likely to be halo-exhibiting raters when rating criterion difficulty is forced to be 0.00.

After anchoring, more than a half of the 11 trained novice Korean raters were fit to the model expectations, as shown in fit statistics in Table 13. When implementing the standard-order analytic rating rubric, raters 03, 06, 09, and 12 provided too predictable ratings. In the reverse-order analytic rating rubric, raters 03 and 06 were overfit to the model expectations. Lastly, when rating criteria were randomly displayed to raters, raters 03, 06, and 12 were classified as muted raters.

All in all, raters 03 and 06 were consistently classified as halo-exhibiting raters regardless of rating criteria order. Rater 09 exhibited a halo in the standard-order analytic rating rubric, and rater 12 exhibited a halo in both standard- and random-order analytic rating rubrics.

Discussion

Halo effect and rating criteria order

The research question posed in this study regarded the extent to which rating criteria order was associated with the magnitude of the group- and individual-level halo effects exhibited by 11 trained novice Korean raters. The results of a three-facet rating scale model suggested that no sizeable group-level halo effect across all three rating criteria orders was detected, as demonstrated in the acceptable fit statistics and halo indices in Table 11. However, the magnitude of the group-level halo effect varied depending on rating criteria order. For example, a similar magnitude of the group-level halo effect was detected in the standard- and reverse-order analytic rating rubrics, while a lesser degree of the group-level halo effect was detected when presenting the four rating criteria in random order.

Although the difficulty measures of the four rating criteria remained invariant in both the standard- and reverse-order analytic rating rubrics, as indicated by 95% confidence intervals in Table 12, *content* and *organization* did not seem to maintain invariance in the random-order analytic rating rubric. More specifically, *content* on average received harsher or lower ratings from the trained novice Korean raters in the random-order analytic rating rubric compared to the reverse-order analytic rating rubric, while the difficulty measure of *content* remained invariant between the standard- and the random-order analytic rating rubrics.

The rating criterion *organization* was more subject to rating criteria order than *content* in this study, as shown in its pronounced difficulty measure in the

Table 13 Rater fit statistics of novice raters after anchoring

Rater	Standard				Reverse				Random			
	Severity	**Infit**	**Outfit**	**Variance**[a]	**Severity**	**Infit**	**Outfit**	**Variance**[a]	**Severity**	**Infit**	**Outfit**	**Variance**[a]
01	0.06	1.01	1.04	0.78	0.44	1.11	1.13	0.89	0.81	0.94	0.94	0.66
02	0.01	1.14	1.17	0.66	0.89	0.84	0.84	0.57	1.86	1.13	1.06	0.54
03	0.47	**0.73**	**0.74**	0.52	0.32	**0.40**	**0.40**	0.36	0.89	**0.69**	**0.69**	0.46
05	0.54	1.33	1.31	0.67	1.27	1.62	1.55	0.54	1.08	1.35	1.35	0.78
06	− 1.57	0.84	**0.79**	0.44	− 0.66	**0.60**	**0.61**	0.31	− 0.62	**0.55**	**0.55**	0.27
07	0.64	1.04	1.04	0.57	0.71	0.89	0.89	0.60	1.04	1.30	1.28	0.81
08	0.00	1.45	1.42	0.87	− 0.53	1.26	1.24	0.58	0.19	0.82	0.82	0.55
09	− 0.04	**0.61**	**0.61**	0.55	1.16	0.86	0.88	0.50	0.34	0.98	0.98	0.73
10	− 0.04	1.10	1.10	1.05	0.59	1.09	1.10	0.94	1.01	1.16	1.16	0.54
11	0.29	0.86	0.88	0.67	1.22	1.40	1.33	0.43	0.54	1.42	1.44	0.81
12	− 0.36	**0.69**	**0.69**	0.43	− 0.08	0.87	0.87	0.48	1.04	**0.54**	**0.54**	0.31

Note: Halo-exhibiting raters whose fit statistics is smaller than 0.80 are in bold

[a] Average within-test taker variances

random-order analytic rating rubric in Table 12. The rating criterion *organization* on average received much harsher or lower ratings from the raters when the raters used the random-order analytic rating rubric than when the standard- and reverse-order analytic rating rubrics were implemented. This finding seems consistent with the previous finding that *organization* was most subject to the halo effect (Lai et al., 2015).

The increased magnitude of the group-level halo effect in the standard- and reverse-order analytic rating rubrics could be attributed to the pre-existing conceptual similarity within two subsets of rating criteria (i.e., *content* and *organization* versus *vocabulary* and *language use*). Furthermore, a successive presentation of two subsets of rating criteria could presumably enhance the pre-existing theoretical similarity from the raters' perspective. In fact, the conceptual similarity is closely related to the assumptions of the rating scale model. As subconstructs of a single latent variable, L2 writing ability, the four rating criteria are supposed to be similar, as implied in the unidimensionality and local independence assumptions of the rating scale model; specifically, those four rating criteria of L2 writing ability should provide "related but independent information" (Andrich & Marais, 2019, p. 173) to estimate L2 writing ability of test taker. However, the framework of communicative language ability predicts that the pair of rating criteria (*content* and *organization*) are conceptually closer to one another than the pair is to the other rating criteria (*vocabulary* and *language use*) as they pertain to textual competence of L2 writing ability (Bachman, 1990). Similarly, the other pair of rating criteria, *vocabulary* and *language use*, is conceptually closer to one another than they are to other rating criteria (*content* and *organization*) as they pertain to grammatical competence of L2 writing ability. Thus, it is plausible to suggest that on top of theoretical similarities among the two rating criteria in each pair, the degree of similarities among the two rating criteria in each subset could be intensified from the raters' perspective as theoretically similar pairs of rating criteria were displayed to the raters in a row.

Relating to the successive presentation of the conceptually similar rating criteria, particularly relevant to the enhanced similarity from the raters' perspective is the spreading activation process in cognitive psychology (Anderson, 2015). The theory is defined as "the process which currently attended items can make associated memories more available" (p. 135). The implication of this theory as an underlying mechanism of the halo effect has been entertained in the literature on the halo effect in rater-mediated performance assessment (Balzer & Sulsky, 1992; Judd et al., 1991; Murphy et al., 1993). Specifically, Balzer and Sulsky (1992) argued that "semantically related dimensions [conceptually similar rating criteria] are thought to be linked in an associative network" (p. 982), emphasizing the need to examine the factors to lessen the degree of the spreading activation process among conceptually similar rating criteria in future research. It could be the case that a rating awarded to the first rating criterion forms an initial impression of the essay, and then the very rating spreads a similar impression on a conceptually similar and immediately following rating criterion from the raters' perspective, thus producing more similar ratings than expected. Thus, the mechanism of the spreading activation process could be applied to explain the following results: when displaying *content* with *organization* to the raters in a row or displaying *vocabulary* with *language use* to the raters in a row as they were in the standard- and reverse-order analytic rating rubrics, the magnitude of the group-level halo effect could become more intense than when presenting rating criteria in random order.

Theoretical and practical implications

The major implications of the current research into the halo effect induced by rating criteria order are threefold. Theoretically, rating criteria order as a common structural design feature in an analytic rating rubric could be incorporated into a framework of L2 writing assessment as it was demonstrated that rating criteria order affected the validity of ratings with respect to the internal structure of rating criteria (i.e., the invariance of rating criterion difficulty) (AERA et al., 2014). The practical implication of the current research is that regarding rubric development, strategies to mitigate the halo effect among the trained novice raters could be considered in light of the importance of rating criteria order. Lastly, the purpose of L2 writing assessment needs to be considered to determine whether the presence of the halo effect poses a threat to the intended use of ratings across rating criteria.

Firstly, rating criteria order should be incorporated into a framework of L2 writing assessment, as the manipulation of rating criteria order in an analytic rating rubric could affect the validity of ratings awarded to test takers. In other words, rating criteria order as a common structural design feature could pose threats to the validity of ratings in rater-mediated L2 writing assessment. Specifically, with respect to the internal structure of the four rating criteria, it was empirically demonstrated that rating criteria order affected “the relationships among test items and test components” (AERA et al., 2014, p. 16) in that the invariance of rating criterion difficulty did not hold depending on rating criteria order, thus introducing either construct-irrelevant easiness or construct-irrelevant difficulty (Messick, 1989). Consequently, it is plausible that the effect of rating criteria order upon the validity of ratings needs to be embraced by a framework of L2 writing assessment.

The practical implication of the current research lies in the concrete recommendations of how to develop analytic rating scales to mitigate the halo effect among newly trained raters when assessing L2 writing ability. As has been suggested by Cooper (1981), the use of improved rating scales is regarded as one of the strategies for reducing the halo effect. In this respect, the results of the current study appear to provide empirical evidence for how to develop analytic rating scales. Specifically, the successive presentation or juxtaposition of conceptually similar rating criteria in an analytic rating rubric might lead to the increased magnitude of the group-level halo effect among newly trained raters. To deal with this issue, the implementation of computer-assisted language assessment (CALT) could make it possible for rating criteria to be randomly presented to newly trained raters to mitigate the halo effect, as the random presentation of rating criteria was found to decrease the group-level halo effect in the current project, as evidenced in halo indices (see Table 11).

When the technology is not available, the rubric format could be manipulated to avoid the successive presentation or juxtaposition of conceptually similar rating criteria.

To illustrate, the implementation of "empirically derived, binary-choice, boundary-definition (EBB) scales" (Upshur & Turner, 1995, p. 6) appears to sidestep this issue; in this rating scale, raters are required to make successive binary decisions for a series of questions designed to capture the ability of L2 writers. Thus, raters would have been unaware of the final score awarded to L2 writers by the time they answer the last question.

The last thing to note is that the magnitude of the group-level halo effect should be interpreted in a manner that facilitates informed decisions about L2 writers. Largely due to the infeasibility of the research design in this study and the unavailability of a large pool of expert raters in most operational settings, it could be difficult to completely dissipate the halo effect. Therefore, the intended use of ratings must be considered when gauging the effects of the halo effect. For example, when the purpose of the assessment is to rank-order L2 writers as for the norm-referenced tests (NRTs), the presence of the halo effect might not affect the intended interpretations of ratings across rating criteria, as the rank-order of L2 writers is, on the whole, unaffected by the flattened rating patterns (Murphy et al., 1993). On the other hand, when the purpose of the assessment is to assess the L2 writing ability "in relation to one or more standards, objectives, or other criteria and not with respect to how much other learners [L2 writers] know" (Carr, 2011, p. 10) in an absolute sense as for the criterion-referenced tests (CRTs), the presence of the halo effect would compromise the intended interpretations of ratings across rating criteria. Furthermore, when the major objective of the assessment is to provide feedback to L2 writers in terms of their strengths and weaknesses, the flattened profile of ratings induced by the halo effect would severely compromise the value of diagnostic feedback to L2 writers (Knoch, 2009).

Conclusion

The results of this study indicated that rating criteria order as the design feature of an analytic rating rubric affected the invariance of rating criterion difficulty in L2 writing assessment as the manipulation of rating criteria order was found to introduce construct-irrelevant easiness or difficulty (Messick, 1989). Thus, rating criteria order as the source of the halo effect could be incorporated into a framework of L2 writing assessment.

The limitations of the current study largely lie in the fact that the findings of this study are restricted to trained Korean raters as the study population. In other words, this study has addressed only the question of how rating criteria order affects the magnitude of the halo effect exhibited by trained Korean raters. In this respect, an important unanswered question in this study involves the degree to which the relationship between rating criteria order in the analytic rating rubric and the magnitude of the halo effect is moderated by the amount of rating experience in operational scoring sessions. Another unanswered question pertains to how the first language of the raters in the study could affect their perceived criterion importance. As the perceived criterion importance leads to distinct rater severity patterns (Eckes, 2012), a commonly shared first language by the trained novice raters in this study could be a moderating variable between rating criteria order and the magnitude of the halo effect.

Appendix

Table 14 An analytic rating rubric

	Content		Organization		Vocabulary		Language use
7	Thorough and logical developments of thesis; substantive and detailed; no irrelevant information; interesting; a substantial number of words for amount of time given	7	Excellent overall organization; clear thesis statement; substantive introduction and conclusion; excellent use of transition word; excellent connections between paragraphs; unity within every paragraph	7	Very sophisticated vocabulary; excellent choice of words with no errors; excellent range of vocabulary; idiomatic and near native-like vocabulary; academic register; no spelling errors	7	No major errors in word order or complex structures; no errors that interfere with comprehension; frequent use of complex sentences; excellent sentence variety
6		6		6		6	
5	Good and logical development of thesis; fairly substantive and detailed; almost no irrelevant information; somewhat interesting; an adequate number of words for the amount of time given	5	Good overall organization; clear thesis statement; good introduction and conclusion; good use of transition words; good connections between paragraphs; unity within most paragraphs	5	Somewhat sophisticated vocabulary; attempts, even if not completely successful, at sophisticated vocabulary; good choice of words with some errors that do not obscure meaning; adequate range of vocabulary but some repetition; approaching academic register; no more than a few spelling errors in less frequent vocabulary	5	Occasional errors in awkward order or complex structures; almost no errors that interfere with comprehension; attempts, even if not completely successful, at a variety of complex structures; frequent use of complex sentences; good sentence variety
4		4		4		4	
3	Some development of thesis; not much substance or details; somewhat uninteresting; limited number of words for the amount of time given	3	Some general coherent organization; minimal thesis statement or main idea; minimal introduction and conclusion; occasional use of transitions words; some disjointed connections between paragraphs; some paragraphs may lack unity	3	Unsophisticated vocabulary; limited word choice with some errors obscuring meaning; repetitive choice of words; no resemblance to academic register; some spelling errors in less frequent and more frequent vocabulary	3	Errors in word order or complex structures; some errors that interfere with comprehension; minimal use of complex sentences; little sentence variety
2		2		2		2	
1	No development of thesis; no substance or details; substantial amount of irrelevant information; very few words for the amount of time given	1	No coherent organization; no thesis statement or main idea; no introduction and conclusion; no use of transition words; disjointed connections between paragraphs; paragraphs lack unity	1	Very simple vocabulary; severe errors in word choice that often obscure meaning; no variety in word choice; no resemblance to academic register several spelling errors even in frequent vocabulary	1	Serious errors in word order or complex structures; frequent errors that interfere with comprehension; almost no attempt at complex sentences; no sentence variety

Acknowledgements
Not applicable

Author's contributions
The author read and approved the final manuscript

Authors' information
Hyunwoo Kim (PhD) is a lecturer in the department of English Language Education in Seoul National University. His research interests include rater bias in rater-mediated performance assessment, and development and validation of rating scales for grammatical complexity.

Funding
Not applicable

Availability of data and materials
The author does not have permission to share data.

Competing interests
The author declares that he has no competing interests.

Received: 29 May 2020 Accepted: 30 October 2020
Published online: 10 November 2020

References

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME] (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Anderson, J. R. (2015). *Cognitive psychology and its implications*, (8th ed.,). New York: Worth publishers.

Andrich, D., Humphry, S. M., & Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement*, *36*(4), 309–324. https://doi.org/10.1177/0146621612441858.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement*, *34*(3), 181. https://doi.org/10.1177/0146621609360202.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. New York: Oxford University Press.

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

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.

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

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110. https://doi.org/10.1191/0265532203lt245oa.

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

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

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780.

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

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: analyzing and evaluating rater-mediated assessments*, (2nd ed.,). Frankfurt: Peter Lang.

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 Retrieved from http://www.jstor.org/stable/1435170.

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). New York: Routledge.

Farrokhi, F., & Esfandiari, R. (2011). A many-facet Rasch model to detect halo effect in three types of Raters. *Theory and Practice in Language Studies*, *1*(11), 1540. https://doi.org/10.4304/tpls.1.11.1531-1540.

Fisicaro, S. A. (1988). A reexamination of the relation between halo error and accuracy. *Journal of Applied Psychology*, *73*(2), 239–244. https://doi.org/10.1037/0021-9010.73.2.239.

Fisicaro, S. A., & Lance, C. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, *14*(4), 419–429.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 241–276). Norwood: Ablex.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher*, *43*(5), 253–263. https://doi.org/10.3102/0013189X14542154.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1*, (pp. 91–118). Kobe: Kobe University.

Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: a practical approach. English composition program*. Rowley: Newbury House.
Judd, C. M., Drake, R. A., Downing, J. W., & Krosnick, J. A. (1991). Some dynamic properties of attitude structures: context-induced response facilitation and polarization. *Journal of Personality and Social Psychology*, *60*(2), 193–202. https://doi.org/10.1037/0022-3514.60.2.193.
Knoch, U. (2009). Diagnostic assessment of writing: a comparison of two rating scales. *Language Testing*, *26*(2), 275–304. https://doi.org/10.1177/0265532208101008.
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. https://doi.org/10.1016/j.asw.2007.04.001.
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. *Educational and Psychological Measurement*, *75*(1), 102–125. https://doi.org/10.1177/0013164414530990.
Lance, C. E., Lapointe, J. A., & Fisicaro, S. A. (1994). Tests of three causal models of halo rater error. *Organizational Behavior and Human Decision Processes*, *57*(1), 83–96. https://doi.org/10.1006/obhd.1994.1005.
Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*, 103–122.
Linacre, J. M. (2018a). *A user's guide to FACETS: Rasch-model computer program*. Chicago: Winsteps.com.
Linacre, J. M. (2018b). *A user's guide to WINSTEPS Rasch-model computer programs*. Chicago: Winsteps.com.
Marais, I., & Andrich, D. (2011). Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, *12*(3), 194–211.
Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective*, (2nd ed.,). New York: Psychology Press.
McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), 52–76. https://doi.org/10.1177/026553229000700105.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). New York: Macmillan.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. https://doi.org/10.3102/0013189X023002013.
Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*(4), 619–624. https://doi.org/10.1037/0021-9010.74.4.619.
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., & 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.
North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(2), 217–263.
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.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465–493. https://doi.org/10.1177/0265532208094273.
Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956–970. https://doi.org/10.1037/0021-9010.85.6.956.
Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12.
Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223. https://doi.org/10.1177/026553229401100206.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287. https://doi.org/10.1177/026553229801500205.
Wells, F. L. (1907). *A statistical study of literary merit: with remarks on some new phrases of the method*. New York: Science Press.
Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: an eye-movement study. *Assessing Writing*, *25*, 38–54. https://doi.org/10.1016/j.asw.2015.05.002.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.