Modern English Education

http://dx.doi.org/10.18095/meeso.2020.21.3.28

Effects of Rating Criteria Order on Novice Korean Raters' Attentional Behaviors in L2 Writing Assessment: A Think-Aloud Protocol Study

Hyunwoo Kim

Seoul National University

ARTICLE INFO

Received: 10 June 2020
Revised: 25 July 2020
Accepted: 14 August 2020

Examples in: English
Applicable Languages: English
Applicable Levels: Tertiary

KEYWORDS

Rater criteria order/
Think-aloud protocol/
L2 writing assessment/
채점순서/
발성사고법/
제2언어 쓰기평가

ABSTRACT

Kim, Hyunwoo. (2020). Effects of rating criteria order on novice Korean raters' attentional behaviors in L2 writing assessment: A think-aloud protocol study. *Modern English Education, 21*(3), 28-36.

Previous studies on rater variability have shown that the amount of attention directed towards each rating criterion was associated with rating criteria order in analytic rating scales. Thus, this study attempts to explore whether rating criteria order affects the amount of attention directed to rating criteria using think-aloud protocols. To achieve a research goal, 11 novice Korean raters rated two essays in two different rating criteria orders. In the standard-order rating rubric, the rating criteria of textual aspects of the essays, including Content and Organization, were first evaluated. Then, the rating criteria of grammatical aspects of the essays, including Vocabulary and Language use, were evaluated. This rating order was precisely reversed in the reverse-order rating rubric. The overall results of this study show that heightened attention from the raters was directed toward either the first- or last-presented rating criterion depending on the quality of each essay. Specifically, a differential amount of attention was directed towards Content and Language Use when rating a poor-quality essay. A significant implication of the study is that rating criteria order should be considered in developing analytic rating scales, and in training newly recruited raters.

I. INTRODUCTION

A think-aloud protocol refers to an introspective method where human participants are asked to concurrently or retrospectively verbalize their cognitive processes as faithfully as possible while engaging in tasks (Gass & Mackey, 2000; Green, 1998). As a sole means to examine cognitive processes of L2 learners, think-aloud protocols have been popular in L2 research (Mackey & Gass, 2015) and in L2 assessment (Bachman & Palmer, 1996). More specifically, verbal reports produced by test takers are believed to present specific validity evidence related to the response process of test takers or raters in the context of L2 assessment.

As a think-aloud protocol renders the cognitive processes of raters directly observable to researchers, a body of research has been extensively conducted to investigate rating processes when raters rate essays in the context of L2 assessment (Barkaoui, 2011; Cumming, 1990; Cumming, Kantor, & Powers, 2002; DeRemer, 1998; Lumley, 2002, 2005; Weigle, 1994; Wolfe, Kao, & Ranney, 1998). The collective findings of the previous studies indicate that previous rating experiences (i.e., expertise) play an essential role to channel raters' impressionistic judgment of the quality of performance into test scores.

In conjunction with think-aloud protocol studies on raters' rating process, eye-tracking studies have shed further light on how a design feature of analytic rating scales is associated with raters' rating process. More specifically, the eye-fixation duration of raters suggests that the posi-

Hyunwoo Kim (Lecturer)
Department of English Language Education, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul, 08826, Korea
Tel: (02) 880-7670 / Email: kim01@snu.ac.kr

tion of a rating criterion in an analytic rating scale is associated with the amount of attention it receives from raters in L2 writing assessment (Winke & H. J. Lim, 2015). Similarly, Ballard (2017) indicated that rating criteria order affects not only the amount of attention from raters, but also interrater reliability and rater severity.

Nonetheless, despite the plethora of think-aloud protocol studies on raters' rating behaviors in L2 writing assessment, few studies have been conducted to examine the extent to which raters' attentional behaviors are altered when manipulating a design feature of analytic rating scales. Thus, in order to occupy this research gap, this study attempts to examine the extent to which rating criteria order in analytic rating scales affects attentional behaviors of trained novice Korean raters as it is suggested that experienced raters might not be subject to the salience of rating criteria induced by the variation of the order of rating criteria in analytic rating scales (Eckes, 2008, 2012). More specifically, based on the results of previous eye-tracking studies (Ballard, 2017; Winke & H. J. Lim, 2015), the researcher in this study implemented two online analytic rating scales where rating criteria order was manipulated, and attempted to examine how the manipulation of rating criteria order alters Korean raters' attentional behaviors, using a concurrent think-aloud protocol.

II. LITERATURE REVIEW

1. Rating Criteria Order

Despite the paucity of think-aloud protocol studies on how rating criteria order affects rating behaviors in L2 writing assessment, two important eye-tracking studies on rating behaviors have shed light on the extent to which rating criteria order is associated with rating process. Winke & H. J. Lim (2015) examined the eye-moment of raters and concluded that the left-most positioned rating criterion in analytic rating scales, *Content* in their study, elicited a heightened amount of attention from raters; on the other hand, the right-most positioned rating criterion, *Mechanics* in their study, received diminished attention from raters.

Following the same research trajectory, Ballard (2017) manipulated the order of rating criteria in analytic rating scales and investigated its effects on the perceptions of 31 inexperienced English-speaking raters, using eye-tracking technology. In the standard-order analytic rating scales, rating criteria are organized in the following chronology: *Content, Organization, Vocabulary, Language use, and Mechanics.* On the other hand, in the reverse-order analytic rating scales, the previous order of rating criteria is precisely reversed. The findings of her study indicated the strong order effects of rating criteria in analytic rating scales. More specifically, raters fixated on the left-most positioned category first, then moved to the right-most positioned category, regardless of the order of rating criteria. Furthermore, raters paid the most attention to the left-most positioned category and the least attention to the right-most positioned category in the standard-order rating rubric, and raters paid a relatively equal amount of attention across categories in the reverse-order analytic rating rubric.

Admittedly, these two studies did not fully address how the order of rating criteria in analytic rating scales affects the rating process of raters, as rating criteria in the analytic rating scales were simultaneously visible to raters; in other words, the influence of a rating awarded to previous rating criteria on ratings awarded to other rating criteria was not successfully controlled. Nonetheless, the empirical results of these two studies demonstrated that 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 & H. J. Lim (2015) indirectly suggest that a certain rating criterion could be artificially rendered salient among other rating criteria by presenting a specific rating criterion first to raters.

2. Think-Aloud Protocol Studies on Rating Behaviors

In the context of L2 assessment, verbal protocol analysis has been regarded as an effective means to examine mental processes of test takers or raters (Green, 1998). To illustrate, this introspective method is implemented to gauge the degree to which similar skills are employed by test takers in completing tasks.

By implementing a concurrent think-aloud protocol, a body of research has been conducted to investigate the differences in rating behaviors between novice and experienced raters. To illustrate, Wolfe et al. (1998) showed that experienced raters focused more on general features of essays (e.g., *Content*) while referring to specific rating criteria in rating scales. On the other hand, novice raters tended to pay heightened attention to specific aspects of essays (e.g., *Language use*), but seldom referred to a rating rubric. In other words, experienced raters appeared to be more versed in terminology in rating criteria originally developed by test developers than novice raters. This finding on the importance of a rating rubric was further corroborated by DeRemer (1998) as it was found that "raters [highly experienced raters] went directly from criterion to criterion following the order of the rubric" (p. 25). Similarly, Cumming (1990) empirically demonstrated that raters displayed dissimilar rating behaviors depending on their previous rating experiences; for example, novice raters largely focused on linguistic aspects of essays, including the classification of errors and the correction of phrases. On the other hand, experienced raters tended to focus more on the content or organization of essays and displayed more self-reflexive behaviors.

Especially relevant to the current study is a descriptive framework outlined by Cumming et al. (2002) showing an explicit delineation of the prototypical rating behaviors. In the last phase of their seminal study, it was found that a similar set of interpretation and judgment strategies were employed by the participants regardless of raters' expertise. Largely due to the robustness of this framework, they

further argue that this descriptive framework should be utilized in "creating checklists of desirable behaviors for raters to use or learn to develop" (p. 88), especially when training newly recruited raters.

3. Purpose

The study population of this study is novice Korean raters who have never rated essays with analytic rating scales. This is mainly because novice Korean 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 (Eckes, 2008, 2012). In other words, in order to isolate raters' beliefs about the relative importance of rating criteria, the population of this study included novice Korean raters whose consistency could be greatly enhanced by receiving rater training, but those who have not yet formulated idiosyncratic beliefs about the relative importance of rating criteria in analytic rating scales.

The independent variable of this study is the variation of rating criteria order in analytic rating scales. More specifically, standard- and reverse-order analytic rating rubrics were implemented online using *Qualtrics*, the web-based survey tool. 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. Additionally, in order to control the direction of the influence of a rating awarded to the preceding rating criteria on ratings on subsequent rating criteria, raters were not allowed to revisit the previous rating criteria to revise scores. In this regard, there exists a huge difference between the independent variable in the previous eye-tracking studies (i.e., the position of rating criteria) and the independent variable in this study (i.e., the presentation order of rating criteria) as the rating criteria were simultaneously displayed to the raters in the previous eye-tracking studies (Ballard, 2017; Winke & H. J. Lim, 2015). On the contrary, the current study forced raters to assess rating criteria in a predetermined order. This constraint was necessary to test the effect of rating criteria order. The dependent variable of this study is the amount of attention directed towards each rating criterion; more specifically, the frequency of segments in verbal reports produced by raters constitutes the dependent variable. The main objective of this study is to answer the following research question:

> To what extent does rating criteria order affect the amount of attention that trained novice Korean raters pay to rating criteria in analytic rating scales?

III. METHODS

1. Participants

The researcher recruited 11 master's students in South Korea whose major was related to Teaching English as Foreign Language (TEFL). The participants, three males and nine females, aged between 26 and 45, reported that they had never rated essays with analytic rating scales. Their proficiency in L2 writing was self-reported as the Common European Framework of Reference for Languages (CEFR) levels, indicating that they were proficient enough to rate the essays using analytic rating scales. More specifically, eight of the participants were classified as C1 (proficient user); two of the participants were as C2 (proficient user); one of the participants was as B2 (independent user).

2. Materials

1) Rating Scales and Essays

The widely implemented analytic rating rubric in the context of ESL writing assessment was modified to answer the research question in the study (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981) (see Appendix). It is important to note a few crucial differences between Jacobs et al.'s analytic rating scales and the revised ones in the current study. Firstly, the rating criterion of *Mechanics*, which was about illegible handwriting, was excluded as the essays had been digitally stored as text files prior to analysis. Additionally, nominal weightings in the original analytic rating scales (e.g., *Content* is given 30 while the rating criterion, *Language use* is given 25) were deleted in order not to distort the importance of rating criteria. Lastly, 25- or 30-point scales of rating criteria collapsed into a 7-point scale to enhance the reliability of ratings given by raters. In the revised analytic rating scales, raters could award 2, 4, and 6 on the 7-point rating scales, whenever they think the defined scales (i.e., 1, 3, 5, and 7 on the 7-point rating scales) do not match the quality of an essay under evaluation.

Two types of 7-point analytic rating scales were implemented using *Qualtrics*, the web-based survey tool: the standard- and reverse-order analytic rating scales. In the standard-order analytic rating scales, rating criteria were displayed to the raters in the following sequence: *Content, Organization, Vocabulary,* and *Language use.* In the reverse-order analytic rating scales, this sequence was precisely reversed. The *Qualtrics* user interface was designed in a way that the raters had to evaluate each essay in a predetermined order and were unable to revisit the previous rating criteria to revise ratings. To illustrate, all rating criteria were not simultaneously displayed to the raters, but only one rating criterion at a time was presented to raters along with the essay under evaluation.

Regarding the subconstruct of second language writing

ability in the study, the revised analytic rating scales were primarily developed to measure organizational competence (Bachman, 1990). Organizational competence is further broken down into grammatical and textual competence; grammatical competence indicates test takers' command of vocabulary, morphology, cohesion, and syntax, while textual competence includes cohesion and rhetorical organization. In the current revised analytic rating rubric, rating criteria including *Content* and *Organization* measure textual competence while rating criteria, including *Vocabulary* and *Language use* measure grammatical competence.

Two argumentative essays on the topic, "smoking in restaurants," were selected from the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2013); the good- and poor-quality essays were selected by the researcher based on the original ratings available in the corpus data, where the same analytic rating scales were implemented (Jacobs et al., 1981). The essays were written by Asian ESL learners and collected for the purpose of building a learner corpus.

2) Coding Scheme

The 11 trained novice Korean raters concurrently produced verbal reports in Korean while rating two essays using the standard- and reverse-order 7-point analytic rating scales. Their verbal reports were transcribed verbatim by the researcher in this study. The main unit of analysis, which is a segment, could be "a phrase, clause or sentence [or sentences]" (Green, 1998, p. 84) as long as it was self-contained enough to facilitate coding decisions without resorting to its adjacent segments. In the current study, the main unit of analysis was decided upon according to whether a segment contained "all information of making an encoding decision" (Ericsson & Simon, 1993, p. 290). To examine the extent to which the rating process of the 11 raters was associated with rating criteria order, a coding scheme was developed and revised in reference to the pre-established descriptive framework of decision-making behaviors (Cumming et al., 2002, p. 88), as shown in Table 1.

TABLE 1
Coding Scheme

Category	Code	Description
Content	A1	Assess thesis and its logical development
	A2	Assess quantity
	A3	Assess interest
Organization	B1	Assess inter-and intra-connection (unity and transition words)
	B2	Assess the adequacy of introduction and conclusion
	B3	Assess thesis statement
Vocabulary	C1	Assess academic register
	C2	Assess sophistication and range of words
	C3	Assess frequency of errors and its gravity in terms of vocabulary
Language use	D1	Assess syntactic complexity
	D2	Assess frequency of errors and its gravity in terms of language use

Note. This table is adapted from Cumming et al. (2002, p. 88)

3. Procedure

1) Rater Training

Before the data collection, the raters participated in two rater training sessions (face-to-face and online rater training sessions), which lasted for about one hour, respectively. Firstly, the face-to-face rater training session was designed to ensure that raters reliably differentiate between rating criteria. Then, raters were asked to participate in the online rater training session, during which they rated three benchmark essays online using the analytic rating scales.

Before conducting concurrent think-aloud verbal protocol analysis, detailed instructions were given to participants to assist them in producing verbal reports, following the guidelines to maximize the validity of verbal protocol analysis (Ericsson & Simon, 1993, p. 377). More specifically, before the verbal protocol session, participants were briefed on what would be required of them throughout the think-aloud protocol study. The researcher provided representative poor-quality samples of verbal reports to sensitize raters to the threats to the validity of concurrent think-aloud verbal protocol analysis. Then, the participants were required to complete two practice tasks in order to familiarize themselves with the technique and to ensure that participants followed the appropriate specified procedures rather than described their general activity.

2) Concurrent Think-Aloud Protocol

Verbal protocol analysis is classified as either talk-aloud or think-aloud depending on the depth of encoding thoughts. More specifically, a talk-aloud protocol asks participants to verbalize the thoughts that are already encoded in verbal form, while a think-aloud protocol requires participants to verbalize even the nonverbal form of thoughts (Ericsson & Simon, 1993). Verbal protocol analysis is also categorized as either concurrent or retrospective protocol, depending on the temporal frame. To illustrate, concurrent verbal reports are produced while human participants are engaged in tasks; on the other hand, retrospective verbal reports are produced after tasks are complete.

It is important to note that think-aloud protocol studies have been criticized due to the veridicality and reactivity of verbal reports produced by human participants (Bowles, 2010). The veridicality of verbal reports is mainly concerned with the accuracy of retrospective verbal reports, as human participants might not be able to accurately reflect their mental processes after tasks are complete. On the other hand, the reactivity of verbal reports is largely concerned with the accuracy of concurrent verbal reports, as the act of verbalizing thoughts itself could alter, if not distort, cognitive processes of raters.

In this study, think-aloud verbal protocol analysis was implemented to capture the thoughts of raters that materialize as either verbal or nonverbal forms while rating the essays. Additionally, to prevent the alteration of information resulting from the rationalization of thoughts after

completing the rating, a concurrent think-aloud verbal protocol analysis was conducted. To minimize the reactivity of verbal reports, the raters were asked to produce verbal reports in Korean, thus precluding the possibility that translation into Korean itself alters the cognitive processes of raters.

After the training session, the raters produced verbal reports in Korean, which were, in turn, audio recorded. The raters were randomly assigned to two groups, as shown in Table 2. More specifically, the rater belonging to Group 01 first produced a verbal report using the standard-order analytic rating scales while rating the first essay; next, the very rater produced a verbal report using the reverse-order analytic rating scales while rating the second essay. In order to counterbalance the order effect of two essays, two essays were also randomly presented to raters. The same sequence and procedure applied to those assigned to Group 02.

TABLE 2
Counterbalancing the Effects of Rating Scales

Rater group	Rating criteria order	
01 (n = 5)	Standard	Reverse
	Good-quality essay	Poor-quality essay
	Poor-quality essay	Good-quality essay
02 (n = 6)	Reverse	Standard
	Good-quality essay	Poor-quality essay
	Poor-quality essay	Good-quality essay

3) Data Coding

After completing the transcription of verbal reports, the researcher in this study and another experienced ESL instructor coded the randomly selected segments, totaling approximately 10% of the entire transcribed segments in order to establish intercoder reliability. Following recommendations by Ericsson and Simon (1993), all the segments were randomly presented to the two coders, thus precluding the possibility that the coders relied on the adjacent segments to make coding decisions. The intercoder reliability was satisfactory, as the simple agreement between two coders reached 92% and the Cohen's κ amounted to .90, which indicates strong agreement between the two coders. Major disagreements arose among A1, B2, and B3 categories; category A1 pertained to the thesis and its logical development while B2 was developed to capture an assessment of the thesis statement. Category B3 was created to evaluate the adequacy of the introduction and conclusion paragraphs. Thus, after discussing the sources of ambiguity with the second coder, the researcher conducted a final coding of those three categories.

4. Exemplary Transcriptions

1) Content

Regarding *Content*, the segment was coded as "assess thesis and its logical development (A1)" when raters assessed the logical development in terms of whether the essay contained substantial details and relevant information to support the main idea (e.g., *It appears that those sentences are not relevant to thesis.* [Rater 10, Segment No. 29]). The segment was coded as "assess quantity (A2)" when raters assessed the word count of the essay as test takers wrote the essay with the limited amount of time (e.g., *The word count was not enough.* [Rater 08, Segment No. 63]). Lastly, the segment was coded as "assess interest" when raters assessed how interesting or genuine the idea was (e.g., *The essay sounds very monotonous.* [Rater 03, Segment No. 260]).

2) Organization

When it comes to *Organization*, the segment was coded as "assess inter- and intra-connectedness (B1)" when raters assessed the unity, coherence of the paragraphs or the use of transition words (e.g., *I don't see any transition words in the body paragraphs.* [Rater 01, Segment No. 54]). The segment was coded as "assess introduction and conclusion (B2)" when raters assessed whether the essay displayed an introduction and conclusion or whether the introduction and conclusion were adequate (e.g., *The writer included only one sentence for the introduction and another sentence for the conclusion. It is very unclear what all reasons in the conclusion referred to in the conclusion.* [Rater 10, Segment No. 11]). The segment was coded as "assess thesis statement (B3)" when raters assessed whether a thesis statement existed and whether a thesis statement contained controlling ideas along with the topic (e.g., *Banning smoking looks like a thesis statement and controlling ideas are cigarettes are harmful, secondhand smoking, and people have the right to be healthy.* [Rater 03, Segment No. 2]).

3) Vocabulary

Regarding *Vocabulary*, the segment was coded as "assess academic register (C1)" when raters assessed whether writers used Latinate verbs instead of their corresponding phrasal verbs or whether the essay exhibited nominalization (e.g., *Substitute, prohibit, and consider, those Latinate verbs, the writer tried to use them.* [Rater 01, Segment No. 12]). The segment was coded as "assess sophistication and range of words (C2)" when raters assessed how sophisticated words were or whether the words were overly repeated (e.g., *The writer should have tried to use other academic verbs.* [Rater 05, Segment No. 19]). The segment was coded as "assess frequency of errors and its gravity in terms of vocabulary (C3)" when raters assessed the errors of spelling and choice of words or how those errors hindered the understanding of the essay (e.g., *Shame should have been shameful.* [Rater 03, Segment No. 210]).

4) Language Use

When it comes to *Language use*, the segment was coded as "assess syntactic complexity (D1)" when raters assessed the syntactic complexity of sentences (e.g., the use

of complex noun phrases) or when raters assessed whether the writer used the variety of sentence structures (e.g., *for smoking is a post-noun modifier.* [Rater 05, Segment No. 34]). The segment was coded as "assess frequency of errors and its gravity in terms of language use (D2)" when raters assessed the frequency of grammatical errors or how those errors hindered the comprehensibility of the essay (e.g., *Affect to their right, no need to use the preposition here.* [Rater 06, Segment No. 60]).

IV. RESULTS

Regarding the good-quality essay, a similar trend regarding all the rating behaviors was observed for both standard- and reverse-order analytic rating scales, as illustrated in the percent of the observations in Table 3.

TABLE 3
Proportion of Group Observations for Two Essays

Coding	Good-quality essay		Poor-quality essay	
	Standard	Reverse	Standard	Reverse
Content	39 (26.4%)	50 (35.0%)	58 (36.9%)	38 (20.2%)
Organization	22 (14.9%)	29 (20.3%)	29 (18.5%)	32 (17.0%)
Vocabulary	22 (14.9%)	26 (18.2%)	35 (22.3%)	43 (23.4%)
Language use	65 (43.9%)	38 (26.6%)	35 (22.3%)	74 (39.4%)
Total	148 (100.0%)	143 (100.0%)	157 (100.0%)	187 (100.0%)

Overall, when the essay was of good quality, a similar amount of attention was directed towards the rating criteria except for *Language use.* It is important to note that the raters seemed to pay an enormous amount of attention to *Content* and *Language use*, regardless of rating criteria order. In contrast, the middle rating criteria, including *Organization* and *Vocabulary* which were the second and third criteria in the rating scales, received scant attention from the raters in both the standard- and reverse-order analytic rating rubrics.

When it comes to the poor-quality essay, a considerable amount of attention was directed toward *Language Use* when raters used the reverse-order analytic rating rubric. Additionally, it is worthwhile to note that when *Content* was displayed first to the raters when rating the poor-quality essay, raters tended to pay a great amount of attention to *Content.* However, when *Content* was displayed last to the raters when rating the poor-quality essay, the attention seemed to diminish.

To further analyze whether rating criteria order was associated with the amount of attention directed towards each rating criterion, a chi-square test of independence was conducted. The statistic tested the null hypothesis that there was no relationship between rating criteria order (i.e., standard- and reverse-order) and the amount of attention directed to each rating criterion at $p < .05$.

As can be seen in Table 4, the results of the chi-square test of independence indicated that there existed a significant association between rating criteria order and the amount of attention directed towards each rating criterion, $\chi^2(3) = 9.65$, $p = .022$, when raters rated the good-quality essay; post hoc comparisons between two rating rubrics across four rating criteria suggested that the amount of attention directed towards *Language use* showed statistically significant difference between two rating scales, $\chi^2(1) = 9.55$, $p = .002$.

TABLE 4
Post Hoc Test for Good-Quality Essay

	Standard	Reverse	χ^2 (*df*)	*p*
Content	39 (26.4%)	50 (35.0%)	2.53 (1)	.112
Organization	22 (14.9%)	29 (20.3%)	1.46 (1)	.226
Vocabulary	22 (14.9%)	26 (18.2%)	0.58 (1)	.447
Language use	65 (43.9%)	38 (26.6%)	9.55 (1)	.002**

Note. Post hoc tests are adjusted for all pairwise comparisons using the Bonferroni correction.
**$p < .01$

As depicted in Table 5, when raters rated the poor-quality essay, there was a significant relationship between rating criteria order and the amount of attention directed towards each rating criterion, $\chi^2(3) = 16.64$, $p = .001$. Post hoc comparisons between two rating rubrics across four rating criteria indicated that *Content* exhibited statistically significant differences in the amount of attention elicited from the raters, $\chi^2(1) = 11.90$, $p < .001$, and that *Language use* in the reverse-order rating rubric elicited more heightened attention than *Language use* in the standard-order rating rubric from the raters, $\chi^2(1) = 11.56$, $p < .001$.

TABLE 5
Post Hoc Test for Poor-Quality Essay

	Standard	Reverse	χ^2 (*df*)	*p*
Content	58 (36.9%)	38 (20.2%)	11.90 (1)	.000***
Organization	29 (18.5%)	32 (17.0%)	0.12 (1)	.726
Vocabulary	35 (22.3%)	43 (23.4%)	0.06 (1)	.810
Language use	35 (22.3%)	74 (39.4%)	11.56 (1)	.000***

Note. Post hoc tests are adjusted for all pairwise comparisons using the Bonferroni correction.
***$p < .001$

Overall, when rating the good-quality essay, there was a significant difference in the amount of attention directed towards *Language use* between the standard- and reverse-order rating rubrics. On the other hand, when rating the poor-quality essay, what was presented first in the rating sequence received the most attention from the raters in the standard- and reverse-order rating rubrics.

V. DISCUSSION AND CONCLUSION

The research question in this study regarded the extent to which rating criteria order is associated with the amount of attention that trained novice Korean raters pay to rating criteria in analytic rating scales. To answer this question, 11 trained novice Korean raters concurrently produced verbal reports while rating two essays written by Asian ESL learners, using the standard- and reverse-order analytic rating scales.

The overall results of the study suggest that the rating

criteria order in the analytic rating rubric appears to be associated with the amount of attention that the raters paid to each of the rating criteria, as evidenced in the significant chi-square test of independence between two rating sequences. Furthermore, the results of subsequent post hoc tests indicated that the rating criterion *Language use* in the standard-order rubric received more heightened attention from the raters than *Language use* in the reverse-order rating rubric when rating the good-quality essay, as evidenced in the statistically significant post hoc comparison. On the other hand, a dissimilar trend was observed when rating the poor-quality essay, as the first presented rating criteria (*Content* and *Language use*) elicited enhanced attention from the raters regardless of rating criteria order; statistically significant post hoc comparisons existed for *Content* and *Language Use* when rating the poor-quality essay.

The results of the quantitative data analysis seem inconsistent with results from the previous eye-tracking studies on rater cognition in L2 writing assessment (Ballard, 2017). With the analysis of raters' total fixation duration on rating criteria, the standard-order rating rubric was found to elicit heightened attention from the raters to the left-most categories (*Content* and *Organization*), but the attention to the right-most categories (*Language use* and *Mechanics*) diminished. On the other hand, the reverse-order rating rubric attracted a similar amount of attention from the raters across the left-most to the middle categories, and slightly heightened attention was directed towards the right-most category (*Content*). This discrepancy could be due to the fact that all the rating criteria were simultaneously visible to the raters in the previous eye-tracking studies, while the current study forced the raters to rate in a predetermined order, precluding the possibility that the raters revised scores on preceding rating criteria.

The expertise of the trained novice Korean raters in this study may have affected the amount of attention paid to each of the rating criteria, as has been suggested in previous studies (Cumming, 1990; Weigle, 1999). Despite the extensive rater training in their research, inexperienced raters tended to focus on different aspects of the essays than did experienced ones and were subject to change in the prompts. As hinted in their studies, 11 trained novice Korean raters could have responded to the alteration of rating criteria order, thus producing qualitatively distinct verbal reports. Additionally, the fact that the first-presented rating criterion received the most attention from the raters might be strongly associated with the impact of fatigue on raters (Ling, Mollaun, & Xi, 2014). In their study, depending on the shift condition, there were significant differences on the rating quality and productivity. In a similar vein, fatigue might have affected the amount of attention paid by the raters in this study.

Additionally, it could be that the quality of the essays might have elicited heightened attention from 11 trained novice Korean raters in this study. Due to the small number of essays in the concurrent think-aloud verbal protocol analysis, it was unclear whether the amount of attention given to the rating criteria was associated with the quality of the essays. To explore this possibility, it would be necessary to examine whether the quality of an essay is confounded with the amount of attention given to rating criteria in future studies.

Admittedly, these arguments cannot be extended to other essays, as only two essays were selected for analysis. Nonetheless, it is suggested that rating criteria order affects the amount of attention from the 11 trained novice Korean raters for these two essays. Furthermore, when rating the poor-quality essay, the first encountered rating criterion elicited enhanced attention from trained novice Korean raters, regardless of rating criteria order. To further confirm this argument, it would be necessary to examine other essays in future studies.

One last aspect to note is that the use of the first language when producing verbal reports can affect the quality and quantity of the verbal reports, as suggested by Barkaoui (2011, p. 19). For example, efforts expended by the raters to translate descriptors or sentences into Korean might have acted as an additional task from the rater's perspective, thus affecting the quality and quantity of the verbal reports. This form of reactivity, the alteration of the rating process induced by the medium of delivery (i.e., language), can pose a great threat to the validity of the concurrent think-aloud protocol analysis study (Bowles, 2010). In this respect, further analysis may be needed to confirm the comparability between the condition with the use of the first language and the condition without the use of the target language in future investigations.

In conclusion, rating criteria order in the analytic rating rubric is associated with the amount of attention that the raters pay to each of the rating criteria. Regardless of rating criteria order, the first encountered rating criterion elicited heightened attention from the raters when rating the poor-quality essay. The rating criterion *Language use* displayed a significant difference in the amount of attention elicited from the raters when rating the good-quality essay. The educational implications of this study might be that rater training sessions should be designed for raters' attention to be equally distributed over each rating criterion of analytic rating scales. Additionally, considering the importance of a rubric in the context of L2 assessment in South Korea (K. R. Lee, 2016), the results of the current study shed light on how to develop a rubric to improve rater consistency in future studies.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: 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). Michigan State University, East Lan-

sing.
Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51.
Bowles, M. A. (2010). *The think-aloud controversy in second language research.* New York: Routledge.
Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.
Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.
DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*(1), 7-29.
Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.
Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*(3), 270-292.
Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.
Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Erlbaum.
Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook* (Vol. 5). Cambridge: Cambridge University Press.
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.* Rowley, MA: Newbury House.
Lee, Kyoung Rang. (2016). The effects of a rubric on inexperienced raters' scoring consistency. *Modern English Education, 17*(2), 75-90.
Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*(4), 479.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276.
Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt: Peter Lang.
Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design* (2nd ed.). New York: Routledge.
Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.
Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145-178.
Winke, P., & Lim, Hyo Jung. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25*, 38-54.
Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication, 15*(4), 465-492.

APPENDIX

Analytic Rating Scales

Score	Content	Organization
7	Thorough and logical developments of thesis; Substantive and detailed; No irrelevant information; Interesting; A substantial number of words for amount of time given	Excellent overall organization; Clear thesis statement; Substantive introduction and conclusion; Excellent use of transition word; Excellent connections between paragraphs; Unity within every paragraph
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	Good overall organization; Clear thesis statement; Good introduction and conclusion; Good use of transition words; Good connections between paragraphs; Unity within most paragraphs
4		
3	Some development of thesis Not much substance or details Somewhat uninteresting Limited number of words for the amount of time given	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
2		
1	No development of thesis; No substance or details; Substantial amount of irrelevant information; Very few words for the amount of time given	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

Score	Vocabulary	Language use
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	No major errors in word order or complex structures; No errors that interfere with comprehension; Frequent use of complex sentences; Excellent sentence variety
6		
5	Somewhat sophisticated vocabulary; Attempts, even if not completely successful, at sophisticated vocabulary; Good choice of words with some errors that don't obscure meaning; Adequate range of vocabulary but some repetition; Approaching academic register; No more than a few spelling errors in less frequent vocabulary	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		
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	Errors in word order or complex structures; Some errors that interfere with comprehension; Minimal use of complex sentences; Little sentence variety
2		
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	Serious errors in word order or complex structures; Frequent errors that interfere with comprehension; Almost no attempt at complex sentences; No sentence variety