# A critical review of "Effects of rating criteria order on the halo effect in L2 writing assessment: A many-facet Rasch measurement analysis"

**Huang Yicai[1] and Zhao Xueai[2]**

[1,2]School of Foreign Studies, Northwestern Polytechnical University, China

**Published: 30 April 2021**


**Abstract**:
*Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis* is one of studies to examine the extent to which the magnitude of the halo effect is associated with rating criteria order in analytic rating scales. This research was conducted by Hyunwoo Kim, a Korean scholar at university of Seoul National University and published in a journal of Language Testing in Asia. This paper aims to summarize the ideas of Kim and make an evaluation of his work.

**Keywords**: Rating criteria order, Halo effect, L2 writing assessment, critical review

## 1. Introduction

The halo effect is defined as "raters' cognitive bias, where the judgement of a certain rating criterion is influenced by that of related other rating criteria of test takers' performance"(Kim, 2020). In the context of rater-mediated performance assessment, a body of studies have examined the impacts of the halo effect on ratings. *Effects of rating criteria order on the halo effect in L2 writing assessment:*

**Cite this article:** Yicai, H. & Xueai, Z. (2021). A critical review of "Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis". *International Journal of Liberal Arts and Social Science*, 9(4), 7-14.

*a many-facet Rasch measurement analysis* is one of studies to examine the extent to which the magnitude of the halo effect is associated with rating criteria order in analytic rating scales. This research was conducted by Hyunwoo Kim, a Korean scholar at university of Seoul National University and published in a journal of *Language Testing in Asia*. This paper aims to summarize the ideas of Kim and make an evaluation of his work.

## 2. Summary

In rater-mediated L2 writing assessment, the halo effect has been proven to be a main source of rater errors. However, the magnitude of the halo effect caused by rating criteria order has not been fully studied. In order to occupy this research gap, Kim's study tried to implement a three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty) to examine the magnitude of the halo effect exhibited by trained Korean novice raters.

During the experiment, firstly, to obtain ratings untainted by the effects of rating criteria order, the researcher recruited four sophisticated ESL instructors at a large state university in the USA to rate 33 essays online. Next, to estimate the magnitude of the halo effect induced by rating criteria order, 11 trained Korean novice raters rated 30 screened essays according to four rating criteria (content, organization, vocabulary, language use) in three different rating orders (standard-, reverse-, and random-order). Then, the researcher analyzed the ratings using a Many-Facet Rasch Model implemented by the software package FACETS. A three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty) was fitted to estimate the group- and individual-level halo effect.

The overall results of Kim's study showed that there was no sizeable group-level halo effect across all three rating criteria order, however, the magnitude of the group-level of halo effect depends on rating criteria order. A similar magnitude of the group-level halo effect was detected in the standard- and reverse-order analytic rating rubrics, and the detected group-level halo effect was lesser when displaying the four rating criteria in random order.

The main implications of Kim's study on the halo effect caused by rating criteria order are threefold. Theoretically speaking, firstly, rating criteria order can be incorporated into the L2 writing assessment framework. Secondly, with regard to rubric development, strategies to reduce the halo effect of trained novice raters can be considered based on the importance of rating criteria order. Finally, the purpose of L2 writing assessment needs to be considered to determine whether the existence of halo effect will threaten the intended use of ratings across rating criteria.

## 3. Evaluation

### 3.1 The interpretation of related literature

Wen (2001) believes that no research is conducted in isolation and it must be linked with the past and future. The literature review not only describes what has been done in the field, but also explains to what extent the researcher's study is different from the previous ones. The main purpose of the literature review is to examine the previous research design, find out the research gap in the previous research, and prove that the current study is worth exploring. However, in my opinion, though Kim

referred to a large number of studies from others, he seldom drew on other studies to support his claims. His study lacked some descriptions of related empirical studies with his evaluating comments.

On the one hand, Kim specified the definitions of seven key terms, but he just listed the conceptual definitions together with references. He neither reviewed critically the various definitions that have surfaced in literature, nor did he explain why he prefer this definition to others. For example, when explaining *validity of ratings*, Kim firstly pointed out that Cronbach & Meehl (1955) divided validity into three types: criterion-oriented validity, content validity, and construct validity. Then, Kim stated that Cronbach & Meehl's classification was criticized by Messick. Messick (1989) suggested that the criterion-oriented validity and content validity should be subsumed under construct validity. Finally, Kim illustrated that *validity of ratings* is defined as the extent to which inferences about L2 writing ability, which is based on Messick's conceptualization. As far as I am concerned, "A, but B, therefore C" is a classical structure in reviewing literature. However, it is obvious that Kim simply enumerated Cronbach & Meehl's and Messick's understanding of validity, but he did not explain why he adopted Messick's conceptualization instead of others. Therefore, I think Kim did not successfully draw on other studies to support his claims.

On the other hand, some terms defined cannot be regarded as key terms and some evaluation of empirical research related to key terms cannot be found. Wen (2001) believes that the variables measured in a study have to be defined conceptually. In Kim's study, obviously, the independent variable is rating criteria order and the dependent variable is the magnitude of halo effect. The instrument for measuring the dependent variable is Many-Facet Measurement Model. Kim wanted to study whether rating criteria order affects the magnitude of halo effect in L2 writing assessment, but he paid more attention to elaborating the definition of validity of ratings, rater bias and rater training. Although these terms are related to the study, they do not belong to the research keywords listed by the author in the abstract. In my mind, Kim is supposed to evaluate some previous empirical research, rather than focus on the definitions of other terms in literature review part. For example, when elaborating relevant empirical research on the halo effect, Kim pointed out that Murphy & Balzer (1989) disconfirmed the hypothesis that the size of the halo effect is negatively associated with some accuracy measures in ratings. He only listed the main content of the empirical research but did not review strengths and weaknesses of this research design. Therefore, I think Kim cited a large number of previous studies, he lacked his own point of view. There is still a lot of room for improvement in literature review.

### 3.2 The implementation of research design

Punch (1998) believes that quantitative research is empirical research where the data are in the form of numbers. In Kim's study, Kim explores the extent to which the magnitude of the halo effect is associated with rating criteria order in analytic rating scales by using Many-Facet Rasch Measurement Model. The data generated through Many-Facet Measurement Model are in the form of numbers, so apparently Kim's study is a typical pure quantitative research. From my point of view, qualitative design is not suitable for Kim's study because the magnitude of the halo effect needs to be measured

by statistical tools. The reasons why the quantitative design used in Kim's study fits for purpose can be discussed from the following three aspects.

First of all, Kim successfully defined variables operationally. In Kim's study, it is apparent that the independent variable is rating criteria order and the dependent variable is the magnitude of halo effect. Kim illustrated the conceptual and operational definition of these two key words in details. The conceptual definition of rating criteria order is that the order in which the criterion appears in the process of rating. The operational definition of rating criteria order is that three types of analytic rating scales (standard-, reverse-, and random-order) are developed using Qualtrics in the process of L2 writing assessment. Halo effect is conceptually defined as raters' undesirable tendency to assign more similar ratings across rating criteria than they should. Operationally, a three-facet rating scale model (L2 writer ability, rater severity, criterion difficulty) is fitted to estimate halo effect. The data of fit statistics of rating criteria are consulted to reflect the magnitude of halo effect. Thus it can be seen that the independent and dependent variables in Kim's study are defined both conceptually and operationally, which makes explicit measurement possible.

Secondly, Kim effectively controlled most of intervening variables. In Kim's study, there are a large number of interference variables, but Kim has successfully controlled most of the interference variables by physical control. In terms of participants, Kim controlled the participants' familiarity with Jacob's analytic rating rubric. Only those who had never rated essays with any variants of the analytic rating rubric originally developed by Jacobs were eligible to participate in the study. In the light of the selected materials, Kim controlled the genre, source and topic of essays. 33 essays are all argumentative essays about smoking in restaurants from the International Corpus Network of Asian Learners of English. According to rater training, Kim took measures to ensure that novice raters could be trained as reliable raters in the study. Novice rates were asked to participate in a face-to-face training. The trained novice raters rated benchmark essays online and ratings provided by them were compared with fair-average scores computed from ratings of the expert raters. If there was a huge difference, the next training would be carried out until reliable ratings were rated. During the implementation of experiment, Kim controlled the time of each rating session. The interval between each rating session was one week. It can be seen that Kim tried his best to avoid the halo effect caused by other factors, which makes the measured data can estimate the extent to which rating criteria order is associated with the magnitude of the halo effect.

Thirdly, Kim made the procedure of experiment systematic, rigorous and scientific. The entire experimental process is mainly divided into two stages. In the first stage, Kim recruited four experienced experts to score 33 selected essays according to the revised analytic rating scales. To obtain ratings untainted by the effects of rating criteria order, only a single rating criterion (content, organization, vocabulary, language use) was rated at a time. The order of rating criteria was counterbalanced with the implementation of a balanced Latin square design. 3 essays were used to test whether the novice raters passed the training and other 30 essays were used for the formal experiment. In the second stage, 12 novice raters participated in both face-to-face and online rating sessions. The raters used the tool *Qualtrics* to rate essays. Three types of analytic rating scales were developed using

*Qualtrics*: the standard-, reverse-, and random-order analytic rating scales. 12 novice raters were randomly assigned to three rating groups and participated in three online rating sessions. To prevent the influence of memories on the results of the experiment, the interval between each rating session was one week. Finally, Kim used FACET software to analyze the collected data to examine the magnitude of group- and individual-level halo effects. Thus it can be seen that the whole experimental process is organized and can achieve the purpose of the experiment.

**3.3 The causality and validity in experimentation**

In the light of the causality in experimentation, Wen (2001) states that three identifying conditions for causality are generally proposed: temporal precedence, necessary connection and the absence of spuriousness within the cause-effect relationship. In my opinion, Kim's empirical research successfully established a causal relationship between the independent variable and the dependent variable. First of all, rating criteria order can affect the magnitude of halo effect. Temporally speaking, rating criteria order evidently occurred before the generation of halo effect. Secondly, the two variables concerned in Kim's study show a necessary link. The changes in rating criteria order are related to the changes in the magnitude of halo effect. Thirdly, in Kim's study, he used the method of physical control to limit intervening variable so that other plausible causes can be ruled out, for example, the rating criterion of mechanics (e.g., illegible handwriting) was excluded in Kim's study. To sum up, the three conditions can be presented simultaneously, so a causal relation is proposed in Kim's study.

In terms of the validity in experimentation, Wen (2001) points out that there two kinds of validity: internal validity and external validity. The internal validity refers to the degree to which the obtained casual relationship can be satisfactorily explained, and the external validity refers to the generalizability of the research findings. Since the generalizability of the findings will be explained in the next section, this section focuses on the internal validity in Kim's study. According to Wen, controlling the intervening variables is a key to success in achieving high internal validity. Only the independent (rating criteria order) and dependent variable (the magnitude of halo effect) are working while the other variables are controlled, the changes in the dependent variable can be attributable to the treatment only. As far as I am concerned, though Kim controlled most of intervening variables, there are still some problems existing in his study. I will illustrate the factors affecting internal validity from three aspects.

Firstly, Kim did not effectively control factors related to the environment. In Kim's study, 4 experts or 12 novice raters all rated essays through online sessions. It was convenient for the researcher to hold online sessions, but environmental factors may affect the results of raters' ratings, such as noise, temperature, time of day, adequacy of light, ventilation, comfort of seats, etc. If the participants were not in the same space, the environmental factors may function together with the independent variable to confound the results.

Secondly, Kim did not effectively control the factors related to subjects. When recruiting experts, Kim tried his best to ensure the similarity of these experts. He hired 4 experienced ESL instructors at a

large state university in the USA. They were quite familiar with Jacobs' analytic rating rubric. When recruiting novice raters, obviously, Kim did not control the gender, age and teaching experience of participants. He hired 12 Korean novice raters with three males and nine females, aged between 26 and 45. They were master's students or in-service teachers and had never rated essays with any variants of Jacobs' analytic rating rubric. In my view, there are apparently more females than male and the age span is extremely large. Also, there is a huge difference in teaching experience between master's students and in-service teachers. These factors related to subjects may function with the independent variable to influence the results.

Thirdly, Kim did not effectively control the factors related to treatments. He recruited 4 experts to rate 33 essays online using the revised analytic rating scales. In order to prevent the raters from remembering their previous ratings for the same article, 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. Similarly, 12 raters were randomly divided into three rating groups and then participated in three online rating sessions accordingly. In order to avoid the effect of memories, 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. From my point of view, the length of the interval of 1 week is not valid. According to Ebbinghaus forgetting curve, 21 days is a person's memory cycle. The interval of 1 week cannot effectively eliminate the influence of memories on the results. So, this factor related to the treatment may function with the independent variable to disorder the results.

### 3.4 The generalizability of the findings

Wen (2001) points out that the external validity refers to the generalizability of the findings. Factors related to the environment and the subjects may affect generalizability. In my opinion, Kim's study can only prove that rating criteria order has impact on the magnitude of the halo effect in L2 writing online assessment. A similar magnitude of the group-level halo effect was detected in the standard- and reverse-order analytic rating rubrics, and the detected group-level halo effect was lesser when displaying the four rating criteria in random order. However, it is enough to extend the findings to a wider context. The reasons can be explained from the following two aspects.

From the perspective of subjects, Bracht & Glass (1968) indicate that the external validity of subjects refers to the extent to which a certain research finding can be applied to other populations. In Kim's study, the participants involved in the experiment were 4 experts and 12 Korean novice raters. The selected materials were 33 argumentative essays from the International Corpus Network of Asian Learners of English. In my view, the number of participants and scoring materials involved in this study is relatively small. When recruiting participants, Kim only considered the participants' familiarity with Jacobs' analytic rating rubric. He did not control factors such as participants' age, gender, teaching experience, ect. When selecting materials and the analytic rating rubric, Kim only took L2 writing assessment into consideration. Therefore, Kim's study can only prove that rating criteria order has impact on the magnitude of the halo effect in L2 writing assessment. It is difficult to extend the findings to other performance assessment.

From the perspective of the environment, Bracht & Glass (1968) point out that the external validity of the environment refers to which a certain research result can be generalized to other settings or contexts. In Kim's study, the participants rated all selected essays through online sessions. They used Qualtrics software to rate the essays in standard-, reverse-, and random-order. As far as I am concerned, Kim strictly controlled the procedures of the experiment, but researchers cannot equate the results of online sessions with the results of offline meetings. The reason is that the online environment is very different from the real life environment. Therefore, Kim's study can only prove that rating criteria order has impact on the magnitude of the halo effect in online writing assessment. It is difficult to extend the findings to offline performance assessment.

## 4. Conclusion: suggestions for further study

Kim's study provides a new perspective for performance assessment and fills the research gap that the magnitude of the halo effect induced by rating criteria order has not been explored. Theoretically speaking, rating criteria order as a common structural design feature in an analytic rating rubric could be incorporated into a framework of L2 writing assessment. The practical implication of the study lies in the concrete recommendations of how to develop analytic rating scales to reduce the halo effect of raters in L2 writing assessment. From my point of view, Kim's study has a complete and clear experimental process, taking into account the operational definitions of independent variables and dependent variables, and controlling interference variables as much as possible. The data analysis is well-founded, and the results obtained are also convincing. However, Kim's study is not a flawless empirical research.

In terms of literature review, though Kim cited a large number of other studies, he seldom explained his own views and lacked a critical evaluation of the previous research design. From the perspective of validity, the internal validity of the research is not high enough, and Kim did not effectively control factors related to the environment, subjects and treatments. With regard to the generalizability of the findings, factors related to the environment and the subjects affect generalizability, which makes the findings unable to be extended to a larger scope. Therefore, I want to give some suggestions from the following aspects. Firstly, when reviewing the literature, researchers should draw previous studies to support their own points of view and to demonstrate the innovation and feasibility of their own research. Secondly, when designing the research, it is necessary to comprehensively consider factors related to the environment, subjects and treatments. Finally, the research findings need to be generalized to a wider context, so researchers are supposed to pay more attention to factors affecting external validity, such as environment and subjects.

## References

[1] Bracht, R.M. & Glass, G.V. (1968). The external validity of experiments. *American Educational Research Journal*, 5(4), 437-74.

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

[3] Kim, Hyunwoo. "Effects of rating criteria order on the halo effect in L2 writing assessment: a many-facet Rasch measurement analysis." Language Testing in Asia 10.1 (2020): 1-23.

[4] Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement, (3rd ed., pp. 13–103). New York: Macmillan.

[5] Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. Journal of Applied Psychology, 74(4), 619–624.

[6] Punch, K. F. (1998). Introduction to social research: Quantitative & qualitative approaches. London: SAGE.

[7] Wen, Q. (2001). Applied Linguistics: Research Methods and Thesis Writing[M]. BeiJing: Foreign Language Teaching and Research Press.