English Language & Literature Teaching, Vol. 19, No. 3 Fall 2013

Student Raters' Reflection on a L2 Writing Binary-Choice Scale in the Context of Self- and Peer-Assessment*

So-yeon Jun**
(Chung-Ang University)
Ho Lee***
(Chung-Ang University)

Jun, So-yeon & Lee, Ho. (2013). Student Raters' Reflection on a L2 Writing Binary-Choice Scale in the Context of Self- and Peer-Assessment. *English Language Literature & Teaching*, *19*(3), 101-123.

This mixed-method study analyzed and revised a binary-choice rating scale for L2 academic writing assessment in order to enlighten how students, exactly pre-service teachers, evaluate a L2 writing binary scale. As the research instrument, Kim's (2010) rating scale, originally developed for Canadian ESL learners, was deployed for analysis and revision. Using data triangulation that included student surveys, student interviews and professional raters' negotiation process, the researcher developed the groundwork for the revision of the Empirically Derived Dichotomous rating scale (EDD). Then, the revision was conducted by the researcher and rater feedback on the revised EDD was collected and analyzed. The findings of the current study indicates that first, for a rating scale to best function in the context of self- and peer-assessment of L2 academic writing, ambiguity that the rater would possibly encounter in the use of the rating scale should be eliminated. Secondly, the revision of a rating scale should be conducted with caution. Even though the majority of the research participants were reportedly satisfied with the revision, some mixed responses to the revision were reported.

[peer assessment/writing assessment/diagnostic test/binary-choice scale]

---

* This study is a part of the first author's M.Ed thesis under the supervision of the corresponding author.
** First author
*** Corresponding author

교보문고 KYOBO Book Centre

I. INTRODUCTION

The previous studies have shown that writing in a second language needs a complicated cognitive process and multi-faced knowledge (Kim, 2010; Lim, 2006; Weigle, 2002). Writing assessment professionals, thus, have tackled with the conundrum on how to properly reflect the spectrum of cognitive features in their rating scale (Weigle, 2002). In solving the challenging task, teachers and researchers as a team elaborate their scale or revise the existing rating scale through pain-staking negotiation process because they believe the selection of a rating scale sometimes determines the quality of the performance assessment.

In the relevant literature, attempts to validate a rating scale have been made by the researchers or testing professionals believed to be eligible for constructing a test. Researchers either have verified reliability and validity of established rating scales adapted to their research contexts (e.g., Barkaoui, 2007, 2008; Hwang, 2008; Kim, 2010; Lee, 2010), or have made tailor-made rating scales for L2 performance assessment (e.g. Hwang, 2008; Kim, 2011; Knoch, 2009; Turner & Upshur, 2002). As aforementioned, efforts have been made to find rating scales that best fit the targeted research or testing context.

Regrettably, however, it was not seriously considered that language learners might provide useful information in the test development process and thereby contribute to the selection of the best rating scale for a given context. The findings of some studies indicated that highly proficient adult learners could be eligible raters of their and their peer's learning outcomes (Lee, 2005, 2007; Lee & Kim, 2009; Lee & Jun, 2011) and that through the involvement in the assessment procedure, further learning can occur (Oh, Shin, & Yager, 2004). As these studies implied, language learners as test consumers are able to play two roles—a test-taker and a rater— in the testing context and provide valuable insights for improving a test.

Another raising concern was that many schools and educational institutions in Korea have adopted rating scales developed by testing companies of English speaking countries without taking the Korean educational context into account (Lee, 2006). For example, Hwang (2008) suggested that the scale developed from English-speaking countries failed to discriminate Korean high school learners and simply provided very limited information to the test taker. According to Yi (2008), the validity of one rating scale could be discussed only in the context where the rating scale was originally developed. That is, the starting point for the validity-relevant arguments of rating should be concerns about the rating scale's adaptability to testing context in operation.

Considering both issues, learners' involvement in rating process and rating scale's adaptability to the specific testing context, this paper will enlighten how the existing

교보문고 KYOBO Book Centre

Empirically Derived Dichotomous (EDD) rating scale, Kim's (2010) rating scale developed in a Canadian ESL context, is modified under the EFL context of student-rater training. Specifically, the EDD is a comprehensive and empirical-based diagnostic kit for writing because a variety of writing criteria in the category of organization, content, and language use were exhaustively collected and empirically tested in the development of the scale. Therefore, the EDD is chosen as the scale to be researched since it is appropriate for the the small-scale and classroom-based assessment.

While the EDD was designed for teachers in need of a diagnostic tool for their students in class (Kim, 2010), the revised EDD, which is the outcome of this study, will be for students who demand a tool for their self- and pee-r assessment of their essay writing. To that end, this study invites pre-service English teachers (and at the same time university students) who were in need of L2 academic writing practice.

Based on the purposes of this research, two research questions are as follows.

1) Among the 35 items of the original EDD, are there any items that need to be revised, added, or deleted? If so, what are the reasons?

2) What are the student raters' perceptions on the revision of the EDD?

II. REVIEW OF LITERATURE

Many rating scale studies have shown that an analytic rating scale could benefit learners because not only it measures the knowledge of the test –taker but also it provides diagnostic information of the test taker (Weigle, 2002). Considering this, a lot of research on the effectiveness of an analytic rating scale has been conducted within the L2 assessment community.

One typical research issue is comparison between a holistic and an analytic scale. The former yields a single score as a result of an overall judgment over a composition, while an analytic rating scale measures several aspects of composition separately (Carr, 2000). Hence, a significant amount of studies aim at finding a better rating scale among the two rating scales for their research and teaching contexts. Many researchers suggested that the analytic rating scale provided much more information to the test users in sacrifice of test practicality (Carr, 2000). However, other research indicated that the holistic measurement showed high inter- and intra reliability (Barkaoui, 2007). In confrontation of such diverse voices, Weigle (2002) suggested that analytic rating scale much better served for diagnostic purposes while the holistic scale was much right for large-scale and high stakes test context.

Another research issue found in the literature on rating scale is the comparison of rating scales that include differential descriptor types. Knoch (2009) made a comparison between

교보문고 KYOBO Book Centre

an intuition-based, experiential rating scale and a data-driven, empirical rating scale. An already established rating scale with less detailed descriptors was compared with an empirically developed one with detailed descriptors. His study found that the latter showed higher reliability and that raters gave more positive feedback on the empirical rating scale. Lee, Kim, and Lee (2011) also buttressed the finding of Knock's study. Lee et al. conducted a comparative study among three kinds of rating scales: general, task-specific, and binary-choice. They found that the empirical-based task-specific rating scale had the highest inter-rater reliability and that the scale was most beneficial to their L2 writing practices.

In recent days, binary-choice rating scales drew positive responses from raters because of their rater-friendly characteristics (E.g., Hwang, 2008; Kim, 2010, 2011; Lee, 2010; Lee, Kim, & Lee, 2011). Facing a binary-choice rating scale, raters in their studies reported that what to look for and what to assess were evident and that a binary choice per a single step was practical and useful for raters. A binary-choice scale forced the rater to choose one of two options, yes or no, for each item in the light of the performance of the test-taker. Shin (2002) stated that it was worth considering the use of dichotomous rating scale in the context of low-stake and diagnostic test.

A binary-choice scale has been frequently used in speaking assessment. One of the most preferred dichotomous scales for L2 speaking assessment is Empirically derived, Binary-choice, Boundary-definition (EBB) rating scale (Upshur & Turner, 1995). The prototypical scale developed by Upshur and Turner (1995) exhibited that every semi-area of the speaking skill had its own tree diagram geared to a specific task. In the tree diagram, a single set of two branches (one branch for 'yes' and the other one for 'no') were arranged hierarchically and a score dangled on every edge of a branch. Raters were required to decide 'yes' and 'no' per each edge in succession up to the final score. In particular, a few studies made a revision to the pioneering model to meet the needs of their Korean EFL subjects (Hwang, 2008; Lee, 2010).

In regard to L2 writing assessment, dichotomous scales were hardly found. Kim's (2010, 2011) studies employed the Empirically-derived Descriptor-based Diagnostic checklist (EDD) in the L2 writing context. Across the rating scale, 35 detailed test-items constituted multiple aspects of L2 writing skill. The rater would check yes if the test-taker's performance meets what the descriptor demands and check no if it does not. What is worth noting was that this descriptor-based assessment tool was empirically derived for the purpose of L2 academic writing unlike other EBB scales designed for L2 speaking rating.

Up to now, the studies on rating scales are reviewed. In the L2 writing area, more precisely in the performance test area, peer-assessment has been gradually paid attention as an alternative assessment. Not to mention the pedagogical effect the peer-assessment have,

교보문고 KYOBO Book Centre

the skepticism of in-classroom performance assessment conducted solely by a teacher have motivated many studies on the peer-assessment within the field of testing. Such research has triggered discussion over the possibility of peer-assessment replacing traditional assessment practices.

Some studies reported the benefits of peer assessment in the writing process. Oh et al. (2004) explored students' perception on peer assessment in an action research context in an attempt to identify pedagogical implications of peer assessment in classroom. The findings of the research indicated that despite negative aspects reported through the students' reflective writing, peer-assessment led to increased ownership of assessment and responsibility of the students. Kim (2005) considered peer-assessment as one way of facilitating learning. She encouraged pre-service teachers to give feedback on their peer's learning outcomes and asked them how they perceived their two roles as a peer feedback giver and a peer feedback receiver. The results of the research revealed that when the participants played as a peer feedback receiver more learning occurred. As for the students' perceptions on the act of giving and receiving peer feedback, the participants reported that peer feedback made them more responsible and cooperative learners.

Some researchers examine the applicability of students' peer-assessment to performance tests in classroom. Kang and Lee (2006) verified the validity and reliability of the students' peer assessment that might resolve the skepticism of the assessment conducted solely by one teacher. The findings of the study showed the possibility of peer-assessment being as an alternative assessment method in classroom. Lee (2005) investigated whether EFL university students' ratings of their peers' L2 writing could be comparable with those of advanced or native graduate students. The finding of the study enlightened that the EFL students' ratings achieved a certain level of validity and reliability when compared with the rating of their counterpart. He argued that as long as the rater could read and comprehend English sentences to some extent, L2 writing proficiency might not be deeply influential in the rating performance of raters. Similarly, Saito and Fujita (2004) explored the acceptance of peer rating in EFL writing classroom. The study found that peer and teacher ratings were correlated significantly with each other.

As shown above, the findings of many studies suggest that an analytic rating scale could provide information that might bring about a positive washback effect. Also, it is believed that a rating scale with detailed descriptors could benefit learners. In regard to the efficacy of peer-assessment, it can be not only reliable and valid but also has valuable pedagogical implications.

교보문고 KYOBO Book Centre

III. METHODS

1. Subjects and Setting

Fifty-three university students were recruited for this study. At the time of recruitment, they majored in English education. The English writing proficiency of the student raters was assumed to be intermediate-high level because the participants scored an average of 23.6 out of 35 points in an essay writing test for class replacement purpose. All student participants played at least three roles in this research: 1) as an essay writer providing essay samples for analysis, three pieces per each, 2) as an essay rater who did a self- and peer-assessment with the provided rating scale and 3) as a questionnaire respondent who fill the survey sheet twice, one before the revision and the other after the revision. Seven students (Seong, Seon, Ji, Mi, In, Hwa, Jin in pseudo-name) recruited from the sample group played an extra role: as an interviewee who was interviewed by the researcher for 30 minutes once before the revision and again after the revision.

The professional rater group consisted of three raters, including the researcher, who had had an experience of English language teaching and English writing assessment, pursuing their master's or doctoral degree in the field of English education. They went through a 2-hour rating workshop where they discussed the EDD rubric and conducted sample ratings. After then, each professional rated 53 writing samples of the student rater with the provided rating scale, the EDD, and participated in the rater conference.

2. Instrumentations: The Original and Revised EDD

The original rating scale that participants referred to in the rating of the first essay was Kim's (2010, 2011) binary-choice rating scale, the EDD. It consisted of thirty five items, representing sub-skills of L2 writing construct. Using the original EDD, both the student rater group and the professional group rated the essays of the student raters. The student raters used the EDD once for their self- and peer-assessment of their essay writing. On the contrary, each professional rater used the EDD 53 times for assessment of the 53 essay samples.

The researcher made a revised version of the EDD, which had thirty items, after collecting students' and expert raters' feedback and the peer researcher's advice. Especially, professional raters' conference made decisions on the direction of the revision and finalized the revised version of EDD. During the revision process, 7 items were kept, 6 items of the original EDD were deleted, 15 items underwent wording changes, 3 items were newly invented, and 6 items were integrated into 3 items. Using the revised EDD the student raters conducted self- and peer-assessment twice with their second and third essay

교보문고 KYOBO Book Centre

writing samples. The interval between their second and third writing tests was one week. Appendix shows the comparison between the original EDD and the revised EDD and the direction of revision per each item.

3. Procedure

1) Writing Tests and Rating Tasks

In each of three essay-writing tests, students were requested to write a four-paragraph English essay on an educational issue within 45 minutes. Each test urged test-takers compose con-and-pro type of essays. After students conducted self- and peer-assessment on their essays, three expert raters rated the writing samples written by the student raters. During the rating, expert raters were solicited to grasp general features of the students' essays and to find recurring features that could not be rated appropriately with the original EDD.

2) Implementation of Open-Ended Survey

After the writing test and rating task were over, open-ended survey written in Korean was administered to the 53-students twice throughout the research procedure, one before the EDD revision (pre survey) and the other after the EDD revision (post survey) as soon as they completed a self- and peer-assessment of their essay. Answering the questionnaire, 53 students were prompted to reflect on the rating scale. The questionnaire sheet includes seven questions regarding rating difficulty, item difficulty and item discrimination, which were constructed to elicit student raters' perceptions on the items of the original and revised EDD. Students were asked to write down any item number that corresponds to each question. Based on the questionnaire answers, seven of the 53 students were invited to participate in the following oral-interview, which was held two weeks after the questionnaire session.

3) Semi-structured Interview

Semi-structured interviews were conducted in Korean to seven students before the revision of the EDD (pre interview) and after the revision of the EDD (post interview). The interviewees participated in a 30-minute individual oral interview with the researcher. Some interview questions were pooled from the questionnaire, while other questions were instantly constructed by the researcher during the interviews to elicit the interviewees' impressions about each item of the rating scale. Interview questions somewhat differed

교보문고 KYOBO Book Centre

from the pre interview and post interview. The pre interviews aimed at checking whether the interviewees had used the original EDD in the intended way, while the post interviews aimed at finding out whether interviewees were satisfied with the revised EDD.

4) Professional Raters' Conference

The researcher and the other two professional raters participated in a rater conference in Korean for making the basis of the revision of the EDD. Prior to this session, the researcher examined the results of the pre survey and pre interviews and identified recurring features of the data, which became sources of the negotiation during the rater conference. Chaired by the researcher, the three professional raters exchanged their reflection on the items of the original EDD and opinions about what should be changed on the current rating scale. Monitoring the rater conference, the researcher mad notes on each item discussed. All talks during the rater conference were audio-recorded and transcribed for research purposes.

4. Data Analysis

This research deploys triangulation of sources which involves using different data sources within the same method in an attempt to guarantee research reliability and credibility. Multiple data sources of survey, interviews and a rater conference are included in this qualitative study. Based on the raters' perceptions on the targeted rating scale, which were collected and analyzed through data triangulation, the current study revised the EDD into a tailor-suited rating scale for the population of the research participants.

For the research question 1, this study utilized the data from pre-survey, pre-interview, and professional raters' conference in order to enlighten the revision process of the EDD. As the research question 2, the researcher included the data from post-survey and post interview for the purpose of gathering the feedback from students, users of the EDD rating scale.

IV. RESULTS AND DISCUSSION

1. Research Question1: The Finding from the Analysis of the Original EDD

1) Student Raters' Survey Analysis

As a result of survey analysis, Table 1 below shows the items with the top two highest frequencies for each of 7 survey questions. Remarkably, items related to content

교보문고 KYOBO Book Centre

fulfillment and vocabulary use gained high frequencies for items that were considered difficult to rate. Meanwhile, items concerning topic-relevance and mechanics were regarded as items with low rating difficulty. As for item difficulty and item discrimination, Item 7 and Item 26 were considered as items with high difficulty and with high discrimination as well. It is assumed that the participants thought difficult items in the EDD can well discriminate their population with writing proficiency. Items concerning article use, vocabulary use and spelling received relevantly many counts as items that need to be revised or deleted.

TABLE 1

Two Most Frequently Mentioned Items of the Original EDD

Survey Items	Most frequently mentioned item	Second most frequently mentioned item
Item with high rating difficulty	Item 7: The supporting ideas and examples in this essay are appropriate and logical. (18)	Item 26: Sophisticated or advanced vocabulary is used. (12)
Item with low rating difficulty	Item 1: This essay answers the question. (25)	Item 34: This essay contains appropriate indentation. (15)
Item with high difficulty	Item 7 (12)	Item 13: Ideas are developed or expanded well throughout each paragraph. (11) Item 26 (11)
Item with low difficulty	Item 1 (23)	Item 33: Capital letters are used appropriately. (14)
Item with high discrimination	Item 7 (21)	Item 26 (12) Item 27: A wide range of vocabulary is used. (12)
Item with low discrimination	Item 33 (18)	Item 1 (17)
Item that needs to be revised	Item 24: Articles are used appropriately. (7) Item 26 (7)	Item 31: Words are spelled correctly. (6)

Note. The numbers in the brackets are the counts of responses that students make.

2) Analysis of Item Interpretations and Reflection of Raters

Twenty-seven items of the EDD were categorized under the four themes, which were: items considered overlapping, items considered unnecessary, items whose wordings need to be amended, and items that need explanation or examples for the raters.

(1) Items considered overlapping

Some student raters pointed out that item 7 (The supporting ideas and examples are appropriate and logical) and item 8 (The supporting ideas and examples are specific and detailed) did not look much different from each other. For instance, Seong said that item 7

교보문고 KYOBO Book Centre

and item 8 were functionally intertwined with each other. Seong mentioned that "if the supporting ideas and examples were appropriate and logical, they would be explained in detail and specific". The professional raters also agreed with Seong's suggestion and decided to merge item 7 and item 8, deleting the words *appropriate* and *detailed* from item 7 and item 8, respectively. Regarding the reasoning behind the deletion of the two adjectives, professional raters asserted in the professional raters' conference that *being appropriate* is a prerequisite to *being logical* while *being detailed* is subsumed to *being specific*.

Some student raters did not differentiate Item 26 (Sophisticated or advanced vocabulary is used) and Item 27 (A wide range of vocabulary is used). Even though Seon reported that "while item 27 is about the abundance of vocabulary in the essay, item 26 more focuses on the difficulty level of vocabulary", she failed to see the notable difference between two items. Another student rater, Yoon, and a professional rater A also reported that item 26 and item 27 seems to measure the same construct.

According to Milton (2009), one's lexical sophistication can be measured by the calculation of the proportion of infrequent words in one's text. There is, however, no convenient way for the student raters to judge whether one word is frequent and others are not, under their self- and peer-assessment context. Taking account of the diagnostic purpose of the learner-directed assessment, the professional raters concluded that item 27 only was sufficient.

(2) Items considered unnecessary

Some student raters pointed out the lack of discriminatory power in item 2 (This essay is written clearly enough to be read without having to guess what the writer is trying to say). In the following Excerpt 1, Seong was concerned about discrimination of item 2. Expert rater B also said that there was no essay that prevented her from comprehending what the writer is trying to say.

> [Excerpt 1]
> Seong: Item 2 has the lowest discriminatory power; no matter how had the quality of the essay is, I could understand what the author is trying to say.
> Researcher: You mean that regardless of the quality of the essay, any essay will be given *yes* for item 2?
> Seong: Yes. As long as the rater has the ability to assess academic essay writing…

In the same vein, item 19 (Grammatical or linguistic errors do not impede comprehension) is not necessary since there existed no single case in which grammatical

교보문고 KYOBO Book Centre

errors in the essay hampered the comprehension of the essay. Accordingly, it was concluded that item 19 was unnecessary in the self- and peer-assessment of this context.

As for the item 29 measuring the ability to employ collocations, the professional raters were skeptical about the usefulness of that item. First, as shown in Excerpt 2, if the use of collocations is not observed in one's essay, that would be a problem. Second, interview with some students showed that they had difficulty in recognizing the proper use of collocation. In addition, rater B raised her concern that the score of item 29 might affect the judgment of item 28 (vocabulary choices are appropriate for conveying the intended meaning). In responses to the concerns, the professional raters concluded that the assessment of collocation use is nearly equivalent to that of acceptable vocabulary choices. Hence, it was concluded that it'd better to have item 29 integrated into item 28.

[Excerpt 2]

Rater A: There were student raters who pointed out item 29 as one considered difficult to use because they said they did not exactly know what collocations are. Also, when asked what if the use of collocation was not observed in a well-written essay, students said that it would be hard to check *no* in the item.

Rater B: The serious problem is that it is ambiguous to define the collocations as the construct to be measured. There are many kinds of collocations which include the combination of noun plus noun, verb plus noun and so on. So we cannot get rid of item 29. Wait, see this [the use of collocations] is a word choice. If one failed to utilize collocations properly, that is the same as he made the wrong word choices.

Rater A: I think there are chances of somebody being confused [over item 29 and item 28].

(3) Items whose wordings need to be amended

Professional raters argued that item 4 (This essay contains a clear thesis statement) did not indicate where a clear thesis statement should occur. As a result of a long discussion, the wording of the item 4 was modified to include the places of thesis statement, in the introductory and conclusion paragraph.

Excerpt 3 shows that when asked for the item 5 (The main arguments of this essay are strong) Ji said that in order for one's arguments to be strong, the evidence presented should be on a solid base. She also stated that what Item 5 was similar to Item 7 (The supporting ideas and examples are appropriate and logical). Professional rater B agreed with Ji's view. Therefore, the professional rater group decided that item 5 should be replaced by an item that asked about whether the writer expressed one clear stance on the topic question.

[Excerpt 3]

Researcher: Can you find difference between 'strong argument' and 'clear argument'?

교보문고 KYOBO Book Centre

Ji: I think so. Clear arguments mean that the writer has made his/her consistent point in the essay, while strong arguments means that the writer presents as many supporting ideas or evidence as possible… I think the intention of item 5 is the same as that of item 7.

(…)

Ji: If an essay's main arguments are strong, the writer put his clear stance on the topic, not sitting on the fence.

Since some students mentioned that the wording of item 14 (Transition devices are used effectively) was vague, the professional raters attempted to change the wording of that item, introducing the concept of cohesion. They concluded that item 14 needs to measure the cohesiveness of all grammatical units such as sentences, clauses, and groups, not superficially checking the transitional devices. Excerpt 4 showed expert raters' talks about the modification of item 14.

[Excerpt 4]

Rater B: It was not easy to use item 14. I think item 14 is related to whether one's essay has cohesion in it. Do subordinates belong to transition devices by the way?

Rater A: If there is a moment of the change in subject throughout an essay, some devices should be needed to make the change seem natural. So if the cohesion guarantees the same function as transition devices…

In particular, an adjective *appropriate* or an adverb *appropriately*, which the descriptions of item 20 (Verb tenses are used appropriate), item 23 (Prepositions are used appropriately), item 24 (Articles are used appropriately), and item 28 (Vocabulary choices are appropriate) commonly included, turned out to be unhelpful in assessing the proper use of specific grammatical features. Since the extent of appropriateness varies from rater to rater, the expert raters suggest that the rating scale include the exact number of errors, as shown in Excerpt 5.

[Excerpt 5]

Rater A: I think we need to turn some words or phrases that give some criteria for judging the mastery of some grammatical features into the word *correctly* or *appropriately*. When we ask student raters of the number of errors that could be acceptable for getting *yes*, their answers varied from rater to rater. So, we need to provide a clear guideline about the number of errors.

Rater B: Or why don't we insert the phrase like this, "certain features are used when necessary".

교보문고 KYOBO Book Centre

(4) Items that need additional explanations or examples for raters

Item 30 (Word forms—noun, verb, adjective, adverb, etc—are used appropriately) asks whether words take the right form of a certain form class which the words belong to. Since the form classes are varied, some student raters were not able to answer an example of each from class. Furthermore, as shown in Excerpt 6, Mi completely misunderstood item 30. Mi considered item 30 as one measuring the general use of grammatical features. Professional raters, acknowledging the lack of student raters' knowledge about correct word form, suggested that some examples of the correct word forms of major form classes should be provided to the student raters.

> [Excerpt 6]
> Now I think that at the time of the assessment I did not understand what item 30 was about. I do not know what demands item 30 ask for. I mean that the phrase *word form* can mean many things. It can be about spelling, or word order, or everything. It can be interpreted as many things.

As discussed above, some student raters can struggle with linguistic terms. Yi's (2008) study found the same problem facing the raters. When having the raters think-aloud during their rating, she found that there were some cases in which the raters did not understand the terms presented in the rating scale being used. Yi argued that if the rater did not understand the descriptors of the rating scale being used, the face validity of the test would be threatened.

While the three professional raters rated the 53 essay samples of the student rater group, it was found that some elements need to be added to the current pool of the items of the original EDD. The professional raters expressed concerns, saying that some features or patterns of the student writings were not properly assessed. As a result, they identified three elements that the future EDD should include. Those three constructs to be measured were 1) control of each paragraph length, 2) control of the whole essay length, and 3) control of overall grammar knowledge.

2. Research Question2: The Findings from the Analysis of the Revised EDD

1) Student Raters' Survey Analysis

The survey analysis suggested that the portion of the positive responses was dominant compared to the negative ones. In detail, positive respondent toward the revision of the EDD accounted for 70 percent of the total respondents (37 out of 53 students), while the

교보문고 KYOBO Book Centre

number of respondents who said they saw no difference between before the revision and after the revision was 16.

Table 2 consists of two most frequently mentioned items of the revised EDD as the answer to each survey item. Regarding the items with high rating difficulty, item 25 received the highest number of counts. Some student raters said that they did not have any idea about the proper width of vocabulary use that deserves *yes* for the item. It seems that the phrase *wide range* in the wording of item 25 might have been confusing to the students. In addition, many student raters remained to have difficulty in measuring a content-fulfillment task, item 8. Concerning item 13, respondents reported that they did not have any idea about the proper proportion of the three sentence structures and that they were forced to use their impression to measure the syntactic variety. In contrast, item 11, item 12, and item 29 were considered ones for students to rate easily. Students stated that it was easy to measure observable sub-skills such as mechanics, spelling, indentation and essay length.

TABLE 2

Two Most Frequently Mentioned Items of the Revised EDD

Survey Items	Most frequently mentioned item	Second most frequently mentioned item
Item with high rating difficulty	Item 25: A wide range of vocabulary and linguistic expressions are used. (14)	Item 8: The supporting ideas and expressions in this essay are specific and appropriate. (11) Item 13: This essay demonstrates syntactic variety, including simple, compound and complex sentence structures. (11)
Item with low rating difficulty	Item 11: Each paragraph consists of more than 3 lines. (19) Item 12: This essay observes the given length. (19)	Item 29: Each paragraph contains an appropriate indentation. (17)
Item with high difficulty	Item 8 (18)	Item 13 (14)
Item with low difficulty	Item 29 (20)	Item 27: Most words are spelled correctly (13) Item 28: All punctuation marks and capital letters are used appropriately. (13)
Item with high discrimination	Item 8 (19)	Item 9: All sentences within each paragraph are cohesively connected. (12)
Item with low discrimination	Item 29 (19)	Item 28 (15)
Item that needs to be revised	Item 24: Less than three types of local error are found throughout the essay (refer to item 17 through item 23). (15)	Item 21: The overall use of article is appropriate. (9)

Note. The numbers in the brackets are the counts of responses that students make.

교보문고 KYOBO Book Centre

Concerning the most challenging area in L2 academic writing, students responded that it was hard to elaborate their thoughts about the essay topic (item 8). Notably, item 8 and item 14 were also pointed out by many student raters as items with high rating difficulty. It seems that in many cases the student raters feel difficulty in making a judgment over the item if they as essay writers had difficulty in fulfilling the task of the item. On the other hand, item 29, item 27, and item 28 assessing the use of mechanic are considered as items with low difficulty.

With regard to discrimination, item 8 discriminates high-level writers from lower proficient writers well. Student raters reported that it was a hard task to persuade readers with the writer's standpoint by presenting supporting examples. Since the task demands the writers' high-mental functioning as well as linguistic competence, writers with lower proficiency may be overburdened to manage the writing task. Item 29 and item 28, however, have the weak discriminatory power because students are thought to hardly make mistakes in the area of mechanics.

Finally, students suggested future directions for development or modification of rating scales. A number of respondents said that item 24 double-penalized the test takers. As for item 21, student raters said that the term, *overall use*, caused confusion, suggesting the explicit number of acceptable errors.

2) Analysis of Rater Responses to the Items of the Revised EDD

The amendment of the wordings of some items drew many positive responses and only a few negative responses from the student rater group. Especially, the elaboration of some items and the insertion of determiners helped eliminate vagueness that some raters had reported as the reason of the difficulty in their usage of the EDD before the revision. According to Ji, for example, some items were integrated to have the decent level of discrimination. Further, the description of some items became detailed and specific. In addition, Seong pointed out the reduced number of the total items, from 35 to 30, reporting that having reduced number of total items would make rater save time and energy.

As illustrated in Excerpt 7, Hwa said that the insertion of determiners in the description of some items was helpful in her rating. This is possibly because determiners such as all and every might have lessened her psychological burden caused from subjectivity in judgment. On the other hand, Jin pointed out the fact that vague phrases still remained even in some items of the revised EDD (*overall use* in item 20 and item 21).

교보문고 KYOBO Book Centre

[Excerpt 7]
Words like *all* and *every* made me examine the writing more thoroughly because no error was allowed as long as words like these were presented in the relevant items. In some cases I reread the writing in order not to make mistakes in my assessment.

Excerpt 8 revealed Mi's positive stance toward item 24 (Less than 3 types of local error are found throughout the essay), explaining that the overall grammatical knowledge of one's essay should be measured. However, Jeong objected to Mi's view and mentioned that "instances of error were already checked through item 17 to item 23. So having item 24 is like we make a minus check twice."

[Excerpt 8]
Item 24 is needed. The reason is that making a local error may be a mistake, but if there are so many mistakes, that could be a problem, decreasing the quality of the whole essay. So I think limiting the number of local error types is not a bad idea.

In addition, Mi liked the subsumption of item 29 (This essay demonstrates facility with appropriate collocations) in the original EDD into item 26 (All vocabulary choices are acceptable for conveying the intended meaning) in the revised EDD since she thought proper collocation might be a part of the appropriate word choice On the other hand, In complained of item 26 with the criticism that the subsumption of an item into another was not persuasive. According to In, "the two items (item 29 in the original EDD and item 26 in the revised EDD) measure two different things."

As for item 14 (Transition devices are effectively used), In gave a positive response. She deeply looked through the organization of the essay because of item 9 (All sentences within each paragraph are cohesively connected) of the revised EDD, while superficially checking the transitional devices with the item 14 of the original EDD. Conversely, Jin criticized that the amendment of item 14 was not reasonable. He argued that item 14 and item 9 were not equivalent, saying that item 9 is related to the content of one's essay, while item 14 is about the use of conjunction.

V. CONCLUSION AND SUGGESTIONS

This qualitative study analyzed and revised a binary-choice rating scale for L2 academic writing assessment for diagnostic purposes. As the research instrument Kim's (2010) rating scale, the EDD, was employed for analysis and revision. In order to identify the item to be revised and to seek the reason for the revision, the researcher used three data collection

교보문고 KYOBO Book Centre

methods: student survey, student interviews, and professional raters' conference. Second, to investigate the student raters' perception on the revised EDD, the researcher conducted a student survey and interviews after the revision of the EDD. The research findings are as follows.

First, a large number of items of the original EDD were revised. Only seven items remained intact, while 28 items went through changes. Referring to the feedback from students' feedback and raters' conference, the researcher deleted six items considered unnecessary, changed the wording of the fifteen items in order to clarify and simplify what the items want to measure, inserted three items, and merged six items into three ones. Students and professional raters suggested a variety of reasons for the modification of the items with highest rating difficulty or those with lowest discrimination.

Secondly, it was revealed that students showed positive attitudes toward the revised EDD. They mentioned that the revised EDD had more decent level of item difficulty and item discrimination than the original EDD. They liked the revised EDD because it cleared vagueness found in the item descriptions of the original EDD and it elaborated its description more understandably. Some students preferred the use of the revised EDD because of the reduced number of total items.

What was examined so far will provide interesting aspects for future studies. First, the fact that the revised EDD helped student raters clearly understand the rating items leads to the increased rating reliability and validity. As put by Lee (2003), any problematic item constructed in an unclear and ambiguous way misleads both test-takers and examiners. Yi (2008) also reported that the raters had difficulty in rating the essays because the rater did not understand terms and wordings of the descriptors of the rating scale being used. Students' or raters' active involvement in the development of a rating scale may relieve the concerns of the researchers.

As for another research implication, this study shows the possibility of incorporating learners' perspectives into the framework of test validity arguments. As stated by Alderson, Clapham, and Wall (1995), a test should be in conjunction with "[learners] learning experiences or their purposes for learning the language, the relationship between how they perceive their language abilities and their performance on the test in question (p. 221)". In order to reflect the needs and demands of test-takers, the current study contained introspective reports from test consumers, including test takers and test examiners of test performance and gathered the variety of ideas from learners as a main source for reconstructing a rating scale. Such a rating scale including voices of students might be sensitive to the context and be friendly toward the test-takers.

Another pedagogical implication for this study is that students' involvement in the construction of rating scale triggers learners' extensive use of the metacognitive strategies in writing. While learners are engaged in the assessment activities, they become keenly

교보문고 KYOBO Book Centre

aware of the assessment criteria. This leads to students' systematic monitoring of their weaknesses and strengths in essay-writing. In particular, the EDD contains more than thirty items, which provides the list of checkpoints required for the mastery of the writing. If students make rating practices with the EDD, they are able to bring telescopic analysis on their own writing in the systematic fashion.

Despite the implications presented above, this study has some limitations. First, the outcome of this study was derived only from perceptions of research participants. That is, this study was heavily dependent on raters' introspective reports and the researcher's qualitative analysis method. Consequently, this study could not draw the conclusion that the revision of the EDD was successful. Even though the majority of the participants were satisfied with the revision, the result was not verified through method triangulation. Future studies need to conduct relevant quantitative analyses to verify the perceived indices of the items of the revised EDD.

Second, theoretical consideration was hardly involved in the process of making amendments to the items of the EDD. Verbal or written accounts collected from the research participants, mainly student raters, were the primary source of the rationale for the analysis and revision of the target scale. Hence, the student raters' lack of background knowledge and expertise can be the target of criticism. Nevertheless, this study decided the student raters as the primary source of feedback because the focus of this study is not to exactly measure writing competence but to construct a rating scale matching for a specific context where students monitor their own writing for diagnostic purposes.

Until now, the current study described how the EDD, initially constructed in Canadian ESL context, was revised to fit for the self- and peer-assessment of the pre-service teachers after accepting students' feedback. Since language testing research have mainly considered the rating scale as a means of assessing test takers for selection purposes in testing contexts, it has been scarcely acknowledged that a well-constructed rating scale used in the diagnostic testing context is a strong impetus to learning. We hope that this study provides a mere step to light the effectiveness of EDD rating scale in the diagnostic context. Furthermore, we expect future studies to delve into the performative role of the rating scale that was incessantly revised in order to reflect the needs of the test consumers.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. NY: Cambridge University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107.

교보문고 KYOBO Book Centre

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes*. Unpublished doctoral dissertation. University of Toronto, Toronto.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Applied Linguistics, 11*(2), 208-241.

Hwang, S. S. (2008). Constructing rating scales for oral performance tests at secondary school based on the EBB model. *English Teaching, 63*(2), 307-331.

Kang, A. N., & Lee, G. M. (2006). A generalizability theory approach to investigating the generalizability of performance assessment using student peer reviews. *Journal of Education Evaluation, 19*(3), 107-121.

Kim, M. J. (2005). Peer assessment as a learning method: the effects of assessor and assessee's roles on metacognition, performance, and motivation. *The Korean Society for Educational Technology, 21*(4), 1-28.

Kim, Y.-H. (2010). *An argument-based inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing*. Unpublished doctoral dissertation. University of Toronto, Toronto.

Kim. Y.-H. (2011). Teachers' perception of using the EDD checklist in ESL academic writing assessment: A qualitative study. *English Language Teaching, 23*(3), 17-39.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275-304.

Lee, H. (2005). Examining the effect of L2 proficiency on a peer-rating procedure in the EFL context. *Foreign Languages Education, 12*(4), 211-234.

Lee, H. (2007). Examining learners' participation in the development of the scoring rubric for EFL essay writing assessment. *English Language Teaching, 19*(4), 279-299.

Lee, H. (2010). Exploring the usefulness of empirically-based binary scale for developing pre-service English teachers' expertise in oral assessment. *The Journal of Korean Teacher Education, 27*(3), 215-237.

Lee, H., & Jun, S. Y. (2011). Engaging pre-service English teachers in the rubric development and the evaluation of a creative English poetry. *English Language Literature & Teaching, 17*(4), 339-356.

Lee, H., Kim, H. Y., & Lee, J. H. (2011). A comparison of three rating scales in an argumentative L2 essay test: General, task-specific, and binary-choice scale. *Korean Journal of English Language and Linguistics, 11*(3), 601-626.

Lee, H. K., & Kim, J. H. (2009). Classroom assessment of English oral proficiency: Focusing on self-, peer-, and teacher assessment. *English Language Teaching, 21*(4), 241-264.

Lee, W. K. (2003). *A guide to English language testing*. Seoul: Moonjin Media.

교보문고 KYOBO Book Centre

Lee, Y. S. (2006). Understanding the basic issues in developing English speaking tests: Test methods and rating scales. *Studies in English Education*, 11(2), 1-18.

Lim, J. W. (2006). Korean EFL writers' composing processes: An exploratory study of college students. *English Language & Literature Teaching, 12*(2), 127-152.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol, UK: Multilingual Matters.

Oh, P. S., Shin, M. K., & Yager, R. E. (2004). Student perceptions of peer assessment in an action research context. *Journal of Korean Earth Science Society, 25*(3), 129-141.

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31-54.

Shin, D. I. (2002). The development of descriptors on scales of L2 productive skills utilized by the many-faceted Rasch model. *English Teaching, 57*(4), 469-499.

Turner, C., & Upshur, J. A. (2002). Rating scales derived from student sampled: Effects of the scales maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70.

Upshur, J. A., & Turner, C. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*(1), 3-12.

Yi, J. Y. (2008). Necessity of developing a rating scale for writing assessment in Korean high schools. *Foreign Languages Education, 15*(3), 261-286.

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

APPENDIX

Comparison between the Original EDD and the Revised EDD

Content/Organization

Before the revision (Original EDD)	After the revision (Revised EDD)	Revision direction
1. This essay answers the questions.	1. This essay deals with the right topic and theme.	clarification
2. This essay is written clearly enough to be read without having to guess what the writer is trying to say.		deletion
3. This essay is concisely written and contains few redundant ideas or linguistic expression.	2. This essay contains few redundant ideas.	simplification
4. This essay contains a clear thesis statement.	5. This essay contains a clear thesis statement in the introductory paragraph. 6. This essay contains a clear concluding statement in the concluding paragraph.	division
5. The main arguments of this essay are strong.	3. One stance is maintained throughout the essay.	clarification

교보문고 KYOBO Book Centre

6. There are enough supporting ideas and examples in this essay.		deletion
7. The supporting ideas and examples in this essay are appropriate and logical.	8. The supporting ideas and examples in this essay are specific and appropriate.	integration
8. The supporting ideas and examples in this essay are specific and detailed.		
9. The ideas are organized into paragraphs and include an intro, body, and conclusion.	10. The ideas are organized into 4 paragraphs and include an introduction, two bodies, and conclusion.	clarification
10. Each body paragraph has a clear topic sentence tied to supporting sentences.	7. Each body paragraph has a clear topic sentence tied to supporting sentences.	keep
11. Each paragraph presents one distinct and unified idea.	4. Each paragraph presents one distinct and unified idea.	keep
12. Each paragraph is connected to the rest of the essay.		deletion
13. Ideas are developed or expanded well throughout each paragraph.		deletion
14. Transition devices are used effectively.	9. All sentences within each paragraph are cohesively connected.	clarification
	11. Each paragraph consists of more than 3 lines (about 45 words).	insertion
	12. This essay observes the given length.	insertion

Grammar

Before the revision (Original EDD)	After the revision (Revised EDD)	Revision direction
15. This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures.	13. This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures.	keep
16. This essay demonstrates an understanding of English word order.	14. This essay demonstrates an understanding of English word order.	keep
17. This essay contains few sentence fragments.	15. This essay is free from sentence fragments.	clarification
18. This essay contains few run-on sentences or comma splices.	16. This essay is free from run-on sentences.	clarification
19. Grammatical or linguistic errors do not impede comprehension.		deletion
20. Verb tenses are used appropriately.	17. All verb tenses are used appropriately.	clarification
21. There is consistent subject-verb agreement.	18. There is consistent subject-verb agreement.	keep
22. Singular and plural nouns are used appropriately.	19. Singular and plural nouns are used appropriately.	clarification
23. Prepositions are used appropriately.	20. The overall use of prepositions is appropriate.	clarification
24. Articles are used appropriately.	21. The overall use of articles is appropriate.	clarification

교보문고 KYOBO Book Centre

25. Pronouns agree with antecedents.	22. Pronouns and demonstratives agree with antecedents.	clarifi-cation
	24. Less than 3 types of local error are found throughout the essay (refer to item 17 through 23)	insertion

Word Use & Mechanics

Before the revision (Original EDD)	After the revision (Revised EDD)	Revision direction
26. Sophisticated or advanced vocabulary is used.		deletion
27. A wide range of vocabulary is used.	25. A wide range of vocabulary and linguistic expressions are used.	clarifi-cation
28. Vocabulary choices are appropriate for conveying the intended message	26. All vocabulary choices are acceptable for conveying the intended meaning.	integration
29. This essay demonstrates facility with appropriate collocations.		
30. Word forms (noun, verb, adjective, adverb, etc) are used appropriately.	23. Correct word forms (noun, verb, adjective, adverb, etc) are used.	clarifi-cation
31. Words are spelled correctly.	27. Most words are spelled correctly.	clarifi-cation
32. Punctuation marks are used appropriately.	28. All punctuation marks and capital letters are used appropriately.	integration
33. Capital letters are used appropriately.		
34. This essay contains appropriate indentation.	29. This essay contains appropriate indentation.	keep
35. Appropriate tone and register are used throughout the essay.	30. Appropriate tone and registr are used throughout the essay.	keep

Examples in: English

Applicable Language: English

Applicable Levels: College

So-yeon Jun
Dept. of English Education
Chung-Ang University
221, Heuk-Seok Dong, Dong-Jak Gu
Seoul, 156-756, South Korea
Tel: 010-6666-9439
Email: absolutejune@gmail.com

교보문고 KYOBO Book Centre

Ho Lee
Dept. of English Education
Chung-Ang University
221, Heuk-Seok Dong, Dong-Jak Gu
Seoul, 156-756, South Korea
Tel: (02)820-5393 / C.P.: 010-2354-4479
Email: holee@cau.ac.kr

Received in July, 2013
Reviewed in August, 2013
Revised version received in September, 2013

KYOBO
교보문고
교보문고 KYOBO Book Centre