Exploring Rater Behaviors During a Writing Assessment Discussion

Susie Kim*
(Michigan State University, USA)
Hee-Kyung Lee
(Yonsei University)


This study explores a discussion held among raters for a writing assessment in a Korean university. It investigates rater behaviors that influence their decision-making processes as revealed in their interaction during discussion. Four raters independently assessed student writing samples using CEFR scales and then held a discussion session to agree on a single score for each sample. Observation and analysis of the rater discussion showed that there were differences in the degree to which individual raters’ initial judgments were reflected in the final decisions and that each rater’s argument style affected the degree. Raters’ personality dynamics, appreciation of student effort, and comprehension of students’ intended meaning were found to be prominent factors that influenced the process of score decisions. These findings have important implications for the use of discussion in performance assessment and for the rating process in general.

Key Words: rater behavior, rater discussion, writing assessment, CEFR

1. INTRODUCTION

In the past, Korea has focused on the instruction of grammar and other receptive skills in English education. Nowadays, however, the importance of productive skills such as speaking and writing has been growing. As a result, tests for speaking and writing are
becoming a necessity for college admission or job applications. Most companies now consider applicants’ speaking ability by interviewing them or requiring scores on English speaking tests such as Oral Proficiency Interview-computer (OPIc) and TOEIC Speaking. In colleges, most of the required English language classes often teach English conversation or writing in English. As performance tests are gaining more importance and popularity, more attention is needed on the rating procedure of those tests.

In writing performance assessments, the process of rating is an important aspect of investigating both test validity and reliability (Lumley, 2002). The conventional holistic rating employs two raters to give a fast, impressionistic judgment for each text, with a third rater option when the two disagree, and then the ratings are summed or averaged for the final, single score (Hamp-Lyons, 1995). Though raters often evaluate the text individually, higher agreement between raters is considered critical for more accurate measurement. This study examines the course of rater discussion, a particular type of rating which involves a socially constructed process. According to Broad (1997), discussion among groups of raters is a hermeneutic model of writing assessment and is defined as “two or more judges working to reach a joint decision on the basis of a writing performance” (p. 134). This study attempts to investigate the nature and patterns of rater behaviors in rater discussion and how they influence score decisions.

2. LITERATURE REVIEW

2.1. The Use of Discussion in Writing Assessments

Rater discussion is one option for establishing appropriate ratings in particular contexts such as rater training, selecting benchmark essays, and modifying grading rubric. In fact, research on rater discussion has been held for those purposes. Studies have investigated the role of rater discussions in assessing portfolios (Broad, 1997; Huot, 1996; Moss, 1996; 1998), resolving discrepant ratings assigned by different raters (Cronbach, Linn, Brennan, & Haertel, 1997; Johnson, Penny, & Gordon, 2001; Johnson, Penny, Fisher & Kuhls, 2003; Johnson, Penny, Gordon, Shumate, & Fisher, 2005), and placing students into composition courses (Huot, 1996). In a study by Moss (1998), group discussion, in which judges collaboratively “engage in dialogue to integrate multiple sources of evidence about a teacher candidate to reach a sound conclusion” (p. 140), was developed and evaluated as a methodology for evaluating portfolios in the teacher licensing process.

Discussion is also used to resolve discrepant independent ratings and settle on a final score. When raters’ initial scores differ, discussion between the original raters or a resolution team composed of expert raters can be used to assign a final score on which all
raters agree (Johnson et al., 2005). Johnson et al. (2005) studied the process of discussion for resolving two discrepant ratings by comparing the discussion method to averaging of the original scores. They investigated whether the raters engaged equally in the resolution process, and the results indicated higher correlations and smaller mean differences between scores reached through discussion and expert-criterion scores, compared with those between averaged scores and expert-criterion scores. The holistic scoring session found possible rater dominance or deference within the group, while the analytic scoring session did not. That is, the final score after discussion in a holistic scoring session agreed more frequently with the original score of one rater, indicating the possibility that one rater tends to be dominant.

2.2. Studies on Rater Variability

Studies on rater effects have mainly focused on rater variability, or differences in the characteristics of raters that lead to inconsistency in the process of rating student writing performance (Eckes, 2008). To handle this problem, rater training has been introduced as a means for maintaining the consistency of assessments. Yet studies have suggested that although intra-rater reliability can be improved by training, rater variables remain a factor (Shohamy, Gordon, & Kraemer, 1992; Weigle, 1998), and training does not have a lasting effect on intra-reliability (Choi, 2002). To date, most major works on rater variability have focused on raters’ differing academic backgrounds and experience (Barkaoui, 2010; Cumming, Kantor, & Powers, 2002; Song & Caruso, 1996; Weigle, 1999; Wolfe, Kao, & Ranney, 1998) or on their language backgrounds (Johnson & Lim, 2009; Shi, 2001).

In addition to raters’ backgrounds, individual differences in cognitive style may have an impact on rater behavior. Baker (2012) argued that individual socio-cognitive characteristics contribute to at least some of the observed rater variability. In this light, some studies identified raters’ cognitive characteristics and their effects on rating. For example, Vaughan (1991) identified five different reading approaches (i.e. “single-focus,” “first impression dominates,” “two-category,” “laughing rater,” and “grammar-oriented rater”) adopted by raters. Milanovic, Saville and Shuhong (1996) categorized four approaches to rating (i.e. “principled two-scan/read,” “pragmatic two-scan/read,” “read through,” and “provisional mark”) from empirical data. In an exploratory study, Baker (2012) regarded score-giving essentially as a decision-making process and explored individual differences in decision-making style. She identified the profile of each rater as “rational,” “intuitive,” “dependent,” “avoidant,” or “spontaneous.” The identification of different styles of raters implies that each rater differs in making judgments.

Past research in rater variability has not fully dealt with the context of discussion, and even the research that has been conducted on rater discussion has not clarified the rating
process of raters in the discussion. This study attempts to address the research gap between these two related topics by observing discussions between raters, which should reveal rater behaviors that may affect rater reliability. The research questions addressed in the study are as follows:

1. To what extent do raters maintain or change their preliminary, independent ratings during the discussion, and what type of argument does each rater adopt?
2. What are some rater behaviors found in group discussion that influence the final score decisions?

### 3. RESEARCH SETTINGS

The context of this study is a South Korean university in which an English placement test is conducted for approximately 3,600 freshmen students each year. The university currently runs a general English program called the English Certification Program, and as the name suggests, its goal is to certify students’ English level. The university adopted the Common European Framework of Reference (CEFR), an internationally accepted set of standards for foreign language curriculum development and assessment, in developing tiered courses for the certification program. In addition, an Internet-based test for English speaking and writing was developed as a placement procedure that places freshmen into one of the university’s English courses according to their proficiency level.

The writing section of the test measures students’ ability to communicate in an academic context. As the curriculum was aligned with CEFR, the placement test also adopted the CEFR scoring rubric. Three “plus levels” for scores on the borderline between two consecutive levels—A1+, B1+, and B2+—were added to the original six CEFR levels; this was done for the purpose of flexible placement in order to adjust the number of students in general English courses. Due to the scope of our study, we focus only on the writing test, which is further described and discussed in the following section.

### 4. METHODS

#### 4.1. Participants

The rater-participants ($N = 4$) were selected to rate a pilot English placement test as a team. They all have thorough knowledge of the CEFR and experiences of teaching various English courses. Table 1 summarizes major characteristics of the raters.
Exploring Rater Behaviors During a Writing Assessment Discussion

### TABLE 1

<table>
<thead>
<tr>
<th></th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>Rater D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>48</td>
<td>51</td>
<td>41</td>
<td>32</td>
</tr>
<tr>
<td>Nationality</td>
<td>Canada</td>
<td>U.S.</td>
<td>Canada</td>
<td>U.K.</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>Male</td>
<td>Male</td>
<td>Male</td>
</tr>
<tr>
<td>Degree</td>
<td>M.A. in TESOL</td>
<td>M.A. in TESOL</td>
<td>M.A. in TESOL</td>
<td>M.A. in History, CELTA</td>
</tr>
<tr>
<td>Years of teaching in Korea</td>
<td>16</td>
<td>17</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>

Note: CELTA = Certificate in Teaching English to Speakers of Other Languages.

Rater A was the team leader. As one of the founding members of the university English placement test, he had participated in developing the placement test items and scoring rubrics, and had trained the raters the previous year. He is from Canada and has a master’s degree in TESOL. He was a certified teacher in Canada and had been teaching English in Korea for 16 years. He was one of the senior members of the general English program and had previously worked as the coordinator of the program.

Rater B is from the United States and has a master’s degree in TESOL. He had been teaching at the university for 10 years and had taught English in other institutions before working at the university. He was also one of the senior members of the program, and the program Coordinator. He had participated in placement test rating since its development.

Rater C is a Canadian with a master’s degree in TESOL. He had been teaching English for 11 years, including teaching at universities in Seoul for seven years. He had been working as a member of the Teaching Development Committee in the program and held the position of Chairman of the College English Tutorial Service Committee at the time of participation in the study. It was his first time rating the university English placement test.

Rater D is from the United Kingdom, holding a Certificate in English Language Teaching to Adults (CELT) issued by the University of Cambridge Local Examinations Syndicate, UK. He had also received IELTS examiner training and had a variety of experience related to teaching and testing English. He had been teaching at the university for three and a half years and had participated in rating the placement test the previous year.

#### 4.2. Data Collection Procedure

##### 4.2.1. Placement test administration

The writing samples used in the rater discussion were collected from a pilot version of the university’s English placement test. This pilot test had several purposes: to collect student samples for rater training, to try out new test items, and to simulate a test situation.
to detect any technical problems that might occur during the actual placement test. The target test-takers were ninety-nine freshmen students who had been enrolled in the general English program for the previous two semesters. The test-takers were sampled proportionally according to the English course distribution: 15% from ESP courses offered to the most advanced learners, 30% from advanced-level courses, 50% from College English, an intermediate level course, and 10-15% from basic-level courses.

### TABLE 2

<table>
<thead>
<tr>
<th>Course Name</th>
<th>Number of Test Takers</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESP</td>
<td>9</td>
<td>9.1</td>
</tr>
<tr>
<td>Advanced College English I (speaking &amp; listening)</td>
<td>14</td>
<td>14.1</td>
</tr>
<tr>
<td>Advanced College English II (reading &amp; writing)</td>
<td>12</td>
<td>12.1</td>
</tr>
<tr>
<td>College English I (speaking &amp; listening)</td>
<td>22</td>
<td>22.2</td>
</tr>
<tr>
<td>College English II (reading &amp; writing)</td>
<td>25</td>
<td>25.3</td>
</tr>
<tr>
<td>Basic College English</td>
<td>13</td>
<td>13.1</td>
</tr>
<tr>
<td>Unidentified</td>
<td>4</td>
<td>4.0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>99</strong></td>
<td>100.0</td>
</tr>
</tbody>
</table>

The placement test was an Internet-based test with an item bank. For each part of the test, one random set of items appeared from the fifteen item sets developed for the part. The writing section consisted of three parts. Part 1 was a warm-up activity which required arranging words in the right order to complete sentences. This section was not taken into consideration when rating. Part 2 involved writing a paragraph about a familiar topic such as “the greatest moment in my life” or “my favorite extracurricular activity.” Examinees were asked to write 100 words in 20 minutes. Part 3 was essay writing on an academic topic such as “corporal punishment” or “evaluating teachers.” This task required 300 words in 35 minutes. Raters used Part 3’s essay task for grading, but if a student did not write this part or if it was too short to provide enough evidence, they referred to the writing from Part 2.

4.2.2. Rating rubric

The rating rubric that the four raters used in the study (see Appendix) consists of five categories (Overall Quality, Language Range, Coherence, Accuracy, and Argument) for nine CEFR levels (A1, A1+, A2, B1, B1+, B2, B2+, C1, C2). The rubric was rearranged according to the publication “Common European framework of reference for languages:
Learning, teaching, assessment (CEFR)” (Council of Europe, 2001). At the outset of the discussion session, the raters agreed to allocate one holistic score to each writing sample according to the lowest marked level among the five categories. This was because the raters agreed that a writing sample cannot be regarded as having achieved one holistic level if it dissatisfies any one of the five criteria.

4.2.3. Rater norming

Norming is a procedure needed to calibrate raters’ understanding of the rubric descriptors. As the rubric was new and one of the raters had no experience in rating for this placement test, it was necessary for the raters to discuss the rubric together to clarify its terms and apply it to several real samples in order to familiarize themselves with it. After the pilot test, the raters used 24 out of 99 sets of student writing samples for norming. The norming session lasted two hours. The raters scored the 24 writing samples independently before the meeting. They then shared their experiences using the rubric to rate this particular set of samples. The raters also discussed how to interpret the CEFR rubric and how to apply the rubric to grading writing samples. This procedure allowed raters to set the rating criteria in their mind in accordance with the CEFR rating rubric.

4.2.4. Rater discussion

Rater discussion was held in order for the four raters to reach a consensus on the writing performance of each student. It should be noted that this discussion was exploited to find benchmark essays representing each CEFR level used for future rater training, and thus the rating process adopted in the study differed from the operational grading of the placement test.

For this procedure, the team members first rated the writing performance samples independently without any discussion. The raters then met to compare and discuss their scores and tried to come to a consensus on each one. The discussion went on for approximately three hours and covered 74 writing samples. Due to the tight schedules associated with preparing for the large-scale placement test, the raters exempted 21 writing samples for which the raters had already assigned the same score in their initial, independent ratings. During the discussion, the raters first shared their independent scores for each sample and then took time to read the sample together in order to recall the thinking processes underlying their preliminary ratings. Then they freely discussed their reasons, relying on the rubric and evidence from the samples. The average time spent on each sample was 3 minutes and 41 seconds.
4.3. Data Coding and Analysis

Data analysis followed the process used in qualitative research by Creswell (2009). Data was prepared for analysis by transcribing the audio-recording of the entire three-hour rater discussion, totaling 16,000 words. The recording was fully transcribed and read several times to obtain a general sense of the information provided and to reflect on the research questions. To address the research questions more effectively, researchers coded the transcript according to two subjects with two subject-specific coding schemes. The two subjects were statements made to maintain or change another rater’s initial rating and characteristics of raters which influenced rating. The agreement rate of the initial coding between the two researchers was .91.

To explore the first research question, analysis included the frequency and percentage of arguments, the samples about which each rater made arguments, and the frequency with which arguments were retained or reflected in the final decision. Prior to coding, researchers consented to the criterion that a statement would be classified as an argument when it contained any of the following five elements: explicit statement of a score, explicit statement of a conventional term for scores, apparent agreement or disagreement in response to another’s opinion, reference to the scoring rubric, and reference to the student writing samples. Examples of arguments are presented in Table 3.

<table>
<thead>
<tr>
<th>Code/Descriptor</th>
<th>Example From Transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit statement of a score</td>
<td>“I think it’s a solid B2”</td>
</tr>
<tr>
<td>Explicit statement of a conventional term for scores</td>
<td>“It was essay-ish” (implies a B1+ level)</td>
</tr>
<tr>
<td>Apparent agreement or disagreement in response to another’s opinion</td>
<td>“I don’t even think it’s that”</td>
</tr>
<tr>
<td>Reference to the scoring rubric</td>
<td>“My point is that ‘occasional unclear expressions and/or inconsistencies may cause break-up in meaning’”</td>
</tr>
<tr>
<td>Reference to the student writing samples</td>
<td>“Look at the second part. How many students do you see who are able to use that construction …and go on with that much accuracy?”</td>
</tr>
</tbody>
</table>

Note. A word or phrase characterizes each score tacitly established among raters.

For the second research question, the researchers reviewed the codes related to rater characteristics by re-evaluating the transcript, and held a discussion to establish final categories and resolve any discrepant coding. This process resulted in three themes which illustrated the rater characteristics that were identified and described both implicitly and explicitly within the data. The results and discussion section includes many excerpts from
rater discussion in order to supplement the description and interpretation of data.

5. RESULTS AND DISCUSSION

The findings of this study reflect the intricate interactions between raters and the influence raters had on one another. Therefore, this section provides excerpts, detailed descriptions, and interpretation. Though data analysis chiefly depended on transcription, the general sense of raters’ attributes gained by listening to the audio recording may have been reflected in the discussion of the findings.

5.1. Retention of Raters’ Initial Rating and Types of Arguments Raters Adopted

This study was interested in to what extent raters maintain or change their initial, independent ratings during the discussion. First, the rate of a rater’s initial score agreeing with the final score was analyzed. Rater C’s independent ratings agreed most often with the final scores: Eighty-two percent of his initially assigned scores matched the final score.

| Table 4 |
| Agreement Between Initial Independent Rating and Final Agreed Score |
|----------------|----------------|----------------|----------------|
| Rater          | Rater A | Rater B | Rater C | Rater D |
| Percentage     | 65%     | 59%     | 82%     | 59%     |

However, the percentages in Table 4 do not provide information about how often the raters expressed their independent opinions in the discussion and how these opinions influenced the final score.

To investigate the extent to which each rater contributed to the discussion, the number of arguments made by each rater was counted. Table 5 shows the number of arguments made by each rater and how many writing samples among the 53 samples each rater made arguments about. When a rater did make an argument for a sample, it was noted whether the statement was retained or withdrawn during the discussion. A rater’s argument(s) were considered as an argument retained when he argued for his initial score and everyone concurred (whether explicitly or not) with his score judgment. When a rater made a statement of agreement regarding another score, it was counted as an argument withdrawn (e.g. “I would agree, it’s not a B2.”). Table 6 shows how often each rater’s arguments remained in effect by indicating the number and percentage of writing samples about which arguments were retained or withdrawn in the final decision.
TABLE 5
The Number of Arguments Made by Each Rater

<table>
<thead>
<tr>
<th>Rater</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>Rater D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of arguments made</td>
<td>47</td>
<td>32</td>
<td>98</td>
<td>42</td>
</tr>
<tr>
<td>Number of samples for which an argument was made</td>
<td>27 (51%)</td>
<td>19 (36%)</td>
<td>42 (79%)</td>
<td>20 (38%)</td>
</tr>
</tbody>
</table>

TABLE 6
The Rate of Argument Retention by Each Rater

<table>
<thead>
<tr>
<th>Rater</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>Rater D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples about which an argument was retained</td>
<td>17 (63%)</td>
<td>13 (68%)</td>
<td>34 (81%)</td>
<td>9 (45%)</td>
</tr>
<tr>
<td>Number of samples about which argument was withdrawn</td>
<td>8 (30%)</td>
<td>4 (21%)</td>
<td>7 (17%)</td>
<td>9 (45%)</td>
</tr>
</tbody>
</table>

Tables 5 and 6 provide insights into the nature of rater arguments and how often each rater’s initial score and arguments were reflected in the final score. In Table 5, the percentage represents the participation rate of each rater. The number of samples about which a rater generated arguments was divided by the total number of writing samples, 53. In Table 6, the percentage provided in the second row indicates raters’ argument retention rate; the number of samples about which an argument was retained in the final score was divided by the total number of samples about which the rater made arguments. And the percentage in the third row means raters’ withdrawal of arguments at the end of the discussions. Rater A made 47 arguments for 27 samples. On average, he made 1.74 arguments per sample. Seventeen of those arguments, or 63%, were reflected in the final score. Following is an example of Rater A’s arguments in discussing one writing sample:

Excerpt 1

Rater A: There are some good things in here but … he just keeps not beginning the sentences with those capitals, not just a few times but it’s a systematic problem.
Rater D: I’m not sure if the density of complex sentences justifies a B2 either. For me, that’s slipping on Range and slipping on Accuracy.
Rater A: And the first thing that they wrote in the very first part, there’s some bad things in there. I mean “My major is architect.” “I listen lecture.” I understand that’s what Koreans say. But “I have honor to France.” What does that mean?
Rater C: Yeah, I’m with you Rater A. B1+.
Rater A: And this here: “teachers do are efforting to adjust more students they can. grading their teachers for student are needed.” That kind of ooh…
Rater C: I agree with you now. It hits the nice format but there are so many other places that have
to go down to a B1+, I think.
Rater B: Alright.

Note. Quotations from the student sample are in double quotation marks, and quotations from the rating rubric are in single quotation marks.

At the outset, Rater A initially marked this sample as B1+, and the other three scored it as B2. First, Rater A points out that the writing has problems with accuracy by saying it has a systematic problem with capital letters. Rater D agrees that it was deficient in accuracy and adds that it failed in range as well. Rater A supports his argument by quoting several sentences from the sample and Rater C responds to this in agreement. Rater A maintained his initial score and successfully convinced the others by quoting from the writing sample to support his argument. Throughout the discussion, he referred to the student writing sample to support his argument, which accounted for 43% of his arguments.

Rater B made arguments about only 19 samples, but his average number of arguments per sample was 1.68, which is similar to Rater A. Thirteen (68%) of his arguments were reflected in the final score. He usually made simple comments to express his opinion. Here are the arguments Rater B generated in discussing one writing sample.

Excerpt 2

Rater B: I think it is a very solid B1+. Let’s see
Rater D: Is it “clear what they’re trying to express”?
Rater B: What they’re trying to say to me is bad students ignore the teacher because there’s no threat. I saw that coming before I read it though.
Rater D: “Occasional unclear expressions”?
Rater B: I would say, with that organization, I absolutely have to give it a B1+…
Rater A: Looking at this, basically I said it was a B1 because of those sentences that I’ve just read to you. So you think that’s not…
Rater B: I got it the first time around. I remember that one.
Rater A: We know what they say, right?
Rater B: Yes.

Raters B and D initially scored this sample as B1+, Rater A as B1, and Rater C as B2. Rater B begins the discussion by saying it is a solid B1+ level essay. Rater D asks a couple of questions to clarify whether the sample has B1+ level accuracy, “Errors occur, but it is clear what s/he is trying to express,” or B1 level accuracy, “Occasional unclear expressions and/or inconsistencies may cause a break-up in meaning.” Rater B insists that the essay is understandable and organized well enough to meet the B1+ level. When Rater A explains
that he gave this sample a lower level because he did not understand several sentences, Rater B states that they were understandable to him even “the first time.” Rater B often made his arguments by merely stating the score or conventional terms referring to scores, which comprises 22% or 25% of his arguments, respectively. He often (28% of his arguments) explicitly argued for a score when the writing sample made sense to him and he had a strong feeling that the sample belonged to a certain level.

Rater C made the most arguments, commenting on 42 samples; this is more than two times the number of samples that Rater B and D commented on. His average number of arguments per sample was also high (2.33), meaning that he not only expressed his opinion about many samples but also he did more often than other raters. Thirty-four (81%) of his arguments were reflected in the final score. Here is an example of Rater C argument about a particular sample.

**Excerpt 3**

Rater C: It was not idiomatic.
Rater B: “lose spirit”
Rater C: It’s not idiomatic, right? I know what she’s saying but it doesn’t have that…
Rater B: The reasons are well supported, it’s argued well.
Rater C: The only thing wasn’t the fluidity of idiomatic language…
Rater D: To be specific?
Rater C: “I thought they love it because they spend same money and less time to move to the store and theater.” I know what’s being said but it’s not idiomatic at all.
Rater B: It’s close.
Rater C: It’s close but it’s not B2+ idiomatic—it’s B2.
Rater B: I think we agree. “quality and quantity of the media attractions. And consumers will…”
That’s a fragment, isn’t it?
Rater C: Just look at the idiomaticity: “Finally, investment for good movie and music will decrease.” Perfectly understandable but….
Rater B: Okay, look at the bottom sentence of that paragraph: “And consumers will lose fun to watch worse movies and listen to worse musics.”
Rater A: Okay, it’s a B2
Rater B: That seals the deal.

In Excerpt 3, Rater C repeatedly argues that the writing sample does not go beyond B2, according to the rubric item “Language lacks expressiveness and idiomaticity.” Other raters do not quite agree that this sample is only a B2 because it is understandable and does not have many grammatical mistakes. When Rater C specifically points out why this
sample lacks idiomaticity with a quote, Rater B sees other expressions that lack idiomaticity, and consequently Rater A agrees that this is a B2 writing sample. Note that Rater C states that this writing sample is “not idiomatic” six times in this excerpt alone. He clearly and persistently stated his opinion. He argued mostly by making references to the scoring rubric, which comprised 40% of his arguments, and often just by stating the conventional term for scores, which comprised 31% of his arguments.

Rater D did not make arguments about many essays, but among those that he did comment on, his average number of arguments per essay was rather high: 2.1 statements. Nine (45%) of his arguments were reflected in the final score. An example of his argument style is as follows.

Excerpt 4

Rater D: That’s “non-systemic.”
Rater C: What’s “non-systemic”?
Rater D: That error, they’re missing the article there. “Is it perfectly same?”
Rater A: I think some of those mistakes are systemic. It’s things like count, non-count nouns, prepositions, uh…
Rater D: I don’t agree that the errors are systemic.
Rater C: What about the idiomaticity?
Rater A: And how about the awareness?
Rater D: I don’t think there’s “occasional slips.” I think there’s nothing fundamentally a fault…
Rater D: I don’t agree that it’s systemic but I agree that the error density pushes it to a certain line.
Rater C: So we’re at B2, we’re good.

In discussing this writing sample, Rater D consistently argued for B2+ while the others argued for B2. The main issue in this writing sample was accuracy. Rater D considers the errors “non-systemic,” which would make this sample a B2+ in terms of accuracy. However, Rater A argues that the errors are systemic, bringing it down to a lower level. While the other raters agree that this sample is a B2, Rater D continues his argument by stating that he does not agree that “the errors are systemic” and does not think there are “occasional slips.” He approves a B2 level in the end due to another issue with the sample, but he still remarks that he does not agree with the opinion that the errors in this sample are systemic. As this excerpt shows, Rater D was sometimes very persistent. He asked many questions throughout the discussion in order to understand the link between a writing sample and the rubric. He tended not to make arguments when he did not have a clear understanding of the link, but when he was sure, he argued forcefully. This tendency explains the relatively lower rate of his arguments being reflected in the final score, as well
as his argument mainly consisting of references to the scoring rubric (55%), or to the writing samples (31%).

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Rater A</th>
<th>Rater B</th>
<th>Rater C</th>
<th>Rater D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit statement of a score</td>
<td>4</td>
<td>9</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>Explicit statement of a conventional term</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>Apparent agreement or disagreement in response to another’s opinion</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Reference to the scoring rubric</td>
<td>19</td>
<td>40</td>
<td>7</td>
<td>22</td>
</tr>
<tr>
<td>Reference to the student writing samples</td>
<td>20</td>
<td>43</td>
<td>9</td>
<td>28</td>
</tr>
</tbody>
</table>

The observation and analyses of raters’ arguments showed their different degree of participation in the discussion. For example, Rater C argued for his score 79% of the time, while Rater B did so only for 36% of the whole discussion. The quality of their arguments could also be inferred from the data. Though Raters B and D argued for their score in a similar number of cases, Rater B was able to retain 68% of his arguments, whereas Rater D retained only 45%. Raters also differed in the types of arguments they adopted, although there is no relationship between the argument type and the argument retention rate.

5.2. Rater Behaviors that Influence Decision Making

5.2.1. Personality dynamics

The arguments and interactions observed in the discussion hinted at the rater’s personalities. In general, Rater A was a rigorous reader with expertise and experience, but he did not force his opinions on others; rather he often used expressions that displayed uncertainty in his arguments and initiated questions to promote discussion. He often encouraged others to participate in the discussion and provided guidance when needed. Rater B assumed a passive attitude during the discussion. He did not speak much and did not like to go against the majority opinion. Toward the end of the rater discussion session, he became even less participatory. Rater C played an assertive and dominant role in the discussion, actively expressing his opinion even when he was in disagreement with all of the others. In most cases, his arguments were well received by the other team members.
Rater D asked a number of clarifying questions and sought concrete evidence to demonstrate that a writing sample belonged to a specific level. He could be very persistent when he had evidence for his argument.

Rater C’s assertive style accounts for the high percentage of agreement between his initial, independent rating and the final score (see Table 4). He spoke very decisively and often succeeded in persuading the rest of the group of his opinion, even when the other three raters had agreed on a different score. Here is an example of an argument that persuaded the other raters.

Excerpt 5

Rater A: Do you think it’s detailed?
Rater C: It’s got enough for me because it gives me examples.
Rater A: Because it does give support. It doesn’t just have topic sentences—it’s got support for them. It’s not a lot but it’s there.
Rater B: At least one support sentence in every paragraph…
Rater D: We need something more definitive. Is there anyone getting anything better on B2, B1+
Accuracy?
Rater A: If you find it, I’ll go for that.
Rater C: I just don’t see that many errors, that’s why. The difference between B1+ and B2 is the amount of errors, right? It’s definitely got errors but it’s not riddled with them.
Rater C: I think there are some complex sentences, too.
Rater D: Yeah I agree. I sort of screened five or six sentences.
Rater B: So B2 it is.
Rater A: See, sometimes the odd man out is actually correct.

The preliminary ratings for this sample were B2 for Rater C and B1+ for the other raters. In this situation, where Rater C had to argue against all of the others, his opinions were accepted. The first five lines illustrate how Rater A and B come to agree with Rater C’s opinion. The rest of the transcript shows acceptance of Rater C’s opinion by Raters A, B, and D. Rater C usually responded quickly with little hesitation, and in a resolute tone. He also had a very firm attitude that dominated the discussion. The following lines suggest his dominance.

Excerpt 6

Rater C: I’m still B1+, You can convince me maybe.
Rater D: I don’t think I can.
The above examples show that the raters found Rater C’s arguments persuasive, valuable, and even unarguable. It can be inferred that his personality had a certain amount of influence on the direction of discussion regarding final score decisions. In contrast, Rater B tended not to argue, especially when his initial rating differed from those of the others. The following examples concern three consecutive samples for which only Rater B had given a different preliminary score from that of the other raters. Rater B did not argue for his score; it was concluded that these three essays belonged to the level initially assigned by the other three raters.

Excerpt 7

Rater B: I’m B1+.
Rater C: So you’re the odd man out.
Rater B: Maybe. (pause) I agree, it’s not a B1+. It’s continuous writing.

Rater B: I said B1+?
Rater C: I don’t think it’s an essay. That was my reason.
Rater B: You’re right. It’s a stream.

Rater C: So I’m at B1.
Rater B: Oh, I got B1+ again, hang on. (pause) Okay, it’s got a start.
Rater C: I just don’t see the essay structure.
Rater B: No, maybe not. (pause) Does it end there?
Rater C: Yep, that’s it.
Rater B: Alright, fair enough.
Rater C: B1?
Rater B: Yep.
Rater A: Hold on, let’s just talk about it for a minute because it has an introduction, and it has a conclusion, and it has a body. The problem with our body is that the body is terrible.
Rater A: See, Rater B, I’m defending you. You won’t step up for yourself so I’m stepping up for you!
Rater B: I’m tired.

Rater B reviews the writing samples again, but he does not make any argument; his refraining may have resulted from various reasons, but the last line clearly shows his disinterest in the discussion. Even when Rater A attempts to persuade Rater B to participate by exploring the possibility that his rating is correct, Rater B avoids arguing, declaring that he is tired.
Before observation of the discussion, it was assumed that the raters’ career backgrounds—such as expertise in rating, previous teaching experience, or current position at the university—would have a major influence on the group decision-making process. However, it was observed that raters’ dispositions had a more significant influence on the decision-making process. Throughout the discussion, Rater C argued most actively and his arguments were the most accepted and reflected in final score decisions (see Table 6), even though he is relatively younger and the least experienced with testing and rating. Rater B, on the other hand, made fewer contributions to the discussion despite his previous experience in the field and the important post he holds in the English program. If he had been more active in the discussion, the decision-making process could have been taken in another direction, as we will see in the last part of this section, because his competence in understanding students’ writing could influence score decisions.

5.2.2. The affective domain

The raters exhibited a characteristic that Turner (2000) called “the affective domain,” that is, recognizing student effort and wanting to somehow work that recognition into making judgments. This study adopted the term to describe the appreciation of student effort by the raters. The following is an example of how the raters took the test situation into consideration.

Excerpt 8

Rater A: … And there are three sentences in there that begin with no capital letters.
Rater D: Oh.
Rater B: Again, it’s a rough draft test situation.
Rater A: But we’ve got to take what they’ve got. I mean that’s what they’ve got. That’s what’s there. I mean you don’t forget to capitalize your sentences.
Rater B: Well I don’t know. I send out plenty of emails, you know, that (are not capitalized). It’s nothing. There’s spell-check that will correct itself later.

In this writing sample, the student did not capitalize some sentences. Rater A considers this an error, but Rater B wants to ignore such mistakes because he understands that students feel pressured when carrying out an unfamiliar writing task in a limited amount of time. Here is a similar example.
Excerpt 9

Rater A: Also, why aren’t they capitalizing stuff they should be capitalizing? It’s weird. They’ve got a fair amount of subject-verb and singular/plural mistakes, as well. …
Rater B: I don’t know if we need to worry about these little capital letter mistakes. That’s just typing. That’s nothing. They are putting together an essay. …
Rater D: I think the capitals are down to the time constraint.

Rater A again disapproves of the omission of capitals. Rater B thinks “these little capital mistakes” are nothing. Rater D also recognizes that it was a test situation in which students were likely to feel pressured due to the time limit. In above occasions, Rater A argues for a lower score due to the capitalization errors while Rater B and D overlook such errors. Below is another example of a different aspect of the affective domain.

Excerpt 10

Rater A: They might actually have thought that their limit was 300 words. And so that’s why they wrote such a short intro and short conclusion. I don’t know. But it’s not a typical hook thesis statement introduction, right? And it’s not giving you much in the end. I mean I think the end is more forgivable than the beginning.
Rater B: I would change. In fact, I would change mine if I was going to do it on my own again, I would give it a +. And I would forgive the first sentence because…
Rater C: Because everything else is good?
Rater B: Well, a lot of times in class I tell them to jump right in. Make your statement and then say “first of all.”

The reason the raters had difficulty assigning a score for this sample is that it had good body paragraphs but one-sentence introduction: “Students should be allowed for several reasons.” Rater A is reluctant to dismiss what he regards as “not a typical hook thesis statement introduction.” Rater B, on the other hand, is willing to accept this short introduction because, as an English instructor, he considers it possible that some students were taught to write that way.

The affective domain seems to have a certain influence on score decisions in terms of how raters embrace the factors that can affect score. Turner (2000) described how this tendency was found in one of the five teachers who were developing rating scales. This study has found a similar behavior in one of the raters and, as the above examples demonstrate, making affective judgments such as allowing students’ pressure and efforts to be reflected in the score may cause leniency in rating.
5.2.3. The level of comprehension of students’ intended meaning

The grading rubric used for the assessment places emphasis on clarity in writing. Descriptors such as “can write clear, detailed text,” “it is clear what s/he is trying to express” and “occasional unclear expressions and/or inconsistencies may cause a break-up in meaning” were frequently mentioned by the raters in their discussion. Consequently, raters who were better at understanding students’ intended meaning tended to give and argue for better grades, and vice versa. Here is an example of a rater changing his decision when his misunderstanding of the text is resolved.

Excerpt 11

Rater A: It’s got a good argument. It’s a good clear argument, definitely. What do you think about the first sentence, “Some people insists that students don’t have to grade their teachers”?
Rater C: They’re saying, some people say you don’t have to grade your teachers but I think you do have to grade your teachers.
Rater A: Ah-ha, you’re right. Why did I give that a B1+? I don’t know why.

This writing sample was an essay in response to the question, “Should students be allowed to grade their teachers?” Rater A assigned a lower score than did Rater C, and he asks for the other’s opinions on a sentence he was unsure about. Rater C shows better comprehension of the sentence in question and explains its meaning. After listening to the explanation, Rater A agrees to give a better score. It can be inferred that the initial misunderstanding did not stem from the student writer’s inability but rather from the rater’s initial inability to comprehend the student’s intended meaning; otherwise, Rater A would not have had the “a-ha” moment and admitted that Rater C’s judgment was correct. Here is another example of raters’ comprehension affecting the decision.

Excerpt 12

Rater B: I think they make the argument case, accuracy… I don’t see any misunderstandings.
Rater C: Yeah, it’s clear enough…
Rater A: For B2, it must be clear and if you look into the first reason, is it clear to you? I don’t really understand what they’re saying there. I understand the sentences but I don’t understand the reasoning I guess. It’s not logical to me.
Rater B: No, no, it is. They drink because they’re curious. They’re curious about this drug, this illegal alcohol. And when they’re freshmen…
Rater C: They want to learn about drinking.
Rater B: Yeah, because…
Rater C: because they’re only doing it because they want to try it.
Rater B: Yeah, when they’re older they’ll be more mature and less curious.

In discussing an essay on the question “Should the legal drinking age be raised?” Rater A argues for the B1+ level, while the others assign B2. Raters B and C rate the sample as B2 because they do not see “any misunderstandings” and it is “clear enough.” However, Rater A complains that some parts of the essay are not clear to him. The discrepancy in scores was caused by the raters’ different levels of comprehension of the students’ intended meaning; when Rater A came to understand, or at least conceded, the logic of the student writer, a resolution was reached.

6. CONCLUSION

This study aimed to increase the understanding of rater behaviors that influence how they reach decisions on final scores during discussion. The frequency and percentage of the arguments made by each rater, and arguments retained and withdrawn, indicated to what extent each rater maintained or changed his preliminary, independent rating during the discussion. Looking more closely into this data also revealed what types of argument each rater preferred to adopt in the discussion. Moreover, this study discovered three potential characteristics that influenced the process of deciding final scores.

Although the nature of this study is exploratory, it provides empirical examples of raters’ characteristics and behaviors in group discussion, which have implications for handling the intricate process of resolving rating discrepancies in writing assessments. First, the findings indicate that rater dominance is possible when scores are determined through rater discussion, as Johnson et al. (2005) has also found. Thus, when rater discussion is used as part of the rating process, it is important to understand the construct-irrelevant factors that can affect the outcome, and to keep the context and the raters’ disposition in mind. In high-stakes tests or other contexts where a precise solution for discrepant ratings is needed, it may be better to employ a third expert rater to handle the discrepancy rather than open discussion among the initial raters.

Rater discussion, however, could be helpful when developing materials for rater training, deciding benchmark essays, or modifying the grading rubric. As raters with different dispositions meet to consider one subject, rater discussion is an amenable platform to air different opinions and problems beforehand and to refine the rating materials and the rating process. In fact, during the discussion, the raters talked about the distinction between the
levels and how certain descriptors were vague or failed to measure students’ ability.

The findings of the present study may be merely suggestive due to its specific context and the small number of raters. In particular, the grading context in the current study might differ from most actual grading contexts in which often a couple of raters are involved in grading each sample and adopts holistic rating rubric. Nevertheless, there exist many occasions in which a group of raters work together with important aims as stated in the above paragraph at a certain stage of performance test development. Considering these contexts, this study can provide useful information on distinctive features of raters working in group discussion. More studies would be required to find differences in decision making process and rating results depending on the number of raters in group discussion. Future studies also need to explore various contexts and diverse groups of raters to accumulate data, through which a clearer picture of rater dynamics and interaction will emerge. In particular, research is needed on which criteria raters rely on most in determining final scores during discussion, as there might be interactions between criteria and raters which were not taken into account in this study.

REFERENCES


## Placement Test Rubric for Written Assessment

<table>
<thead>
<tr>
<th>Overall</th>
<th>Range</th>
<th>Coherence</th>
<th>Accuracy</th>
<th>Argument</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>Can write clear, smoothly flowing, complex texts in an appropriate and effective style and logical structure which helps the reader to find significant points.</td>
<td>Shows great flexibility in utilizing linguistic forms to convey ideas clearly; gives emphasis to primary concern and avoids ambiguity. Shows a good command of idiomatic expressions and colloquialisms.</td>
<td>Can create coherent and cohesive texts making full and appropriate use of a variety of organizational patterns and a wide range of connectors and other cohesive devices.</td>
<td>Maintains consistent and highly accurate grammatical control of even the most complex language forms. Errors are rare and consistently corrected.</td>
</tr>
<tr>
<td>C1</td>
<td>Can write clear, well-structured essays of complex subjects, underlining the relevant salient ideas, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, rounding off with an appropriate conclusion.</td>
<td>Has a good command of a broad range of language allowing him/her to select a formalization to express him/herself clearly in an appropriate style on a wide range of academic topics without having to restrict what he/she wants to say. The flexibility in style and tone is somewhat limited.</td>
<td>Can produce clear, smoothly-flowing, well-structured text, showing controlled use of organizational patterns, connectors and cohesive devices.</td>
<td>Continuously maintains a high degree of grammatical accuracy; occasional errors in grammar, collocations and idioms.</td>
</tr>
<tr>
<td>B2+</td>
<td>Can use a range of language to express abstract ideas as well as topical subjects, correctly most of the mistakes in the process.</td>
<td>Can express him/herself clearly without much sign of having to restrict what he/she wants to say. Can use a variety of linking words to clarify the relationships between ideas.</td>
<td>Good grammatical control; occasional “slip” or non-systemic errors are minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.</td>
<td>Can write an essay that develops an argument systematically with appropriate highlighting of some significant points and relevant supporting detail. Can give some reasons in support of a particular point of view and explain the advantages and disadvantages of various options.</td>
</tr>
<tr>
<td>B2</td>
<td>Can write a detailed essay on a variety of subjects related to his field of interest, synthesizing and evaluating information and arguments from a number of sources.</td>
<td>Has a sufficient range of language to be able to express viewpoints on most general topics, using some complex sentence forms is necessary. Language lacks, however, expressiveness and idiom and use of more complex forms is still stilted. Can make a distinction between formal and informal language with occasional less appropriate expressions.</td>
<td>Can use a number of cohesive devices to link his/her sentences into clear, coherent text, though there may be some “jumpiness” in longer text.</td>
<td>Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstandings.</td>
</tr>
<tr>
<td>B1+</td>
<td>Can write about a variety of familiar subjects well enough for others to follow the story or argument.</td>
<td>Can describe a coherent situation, explaining the main points in an idea or a problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.</td>
<td>N/A</td>
<td>Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable native tongue influence. Errors occur, but it is clear what she is trying to express.</td>
</tr>
<tr>
<td>B1</td>
<td>Can write straightforward connected essays on a range of familiar subjects within his field of interest, linking a series of short discrete elements into a linear sequence.</td>
<td>Has enough language to get by, with sufficient vocabulary to express him/herself with some schematization on topics such as family, hobbies and interests, social events, current affairs.</td>
<td>Can link a series of shorter discrete elements into a connected, linear text.</td>
<td>Can produce continuous writing which is generally intelligible throughout.</td>
</tr>
<tr>
<td>A2</td>
<td>Can write a series of simple sentences with relatively few, if any, adverbs or adjectives.</td>
<td>Uses basic sentence patterns with limited information in everyday situations.</td>
<td>Can link groups of words with simple connectors like “and”, “but” and “therefore”. Use simple structures correctly, but still occasionally makes basic errors. Longer texts may contain expressions and show coherence problems which make them hard to understand.</td>
<td>N/A</td>
</tr>
<tr>
<td>A1+</td>
<td>Can write simple sentences about oneself, for example where one lives and what one does.</td>
<td>If the student meets all the criteria for A2, the overall descriptor of the “A1+” is 2 out of 4 of the descriptors for A2, then the student is at an “A1” level.</td>
<td>Show only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire. Longer texts contain expressions and show coherence problems which make them very hard or impossible to understand.</td>
<td>N/A</td>
</tr>
<tr>
<td>A1</td>
<td>Can write simple isolated phrases and sentences.</td>
<td>Uses a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.</td>
<td>Can link words or groups of words with very basic linear connectors like “and” and “then.”</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Applicable levels: Tertiary

Susie Kim
Second Language Studies Program
Colleges of Arts & Letters
Michigan State University
619 Cedar Road
East Lansing, MI, 48824
United States of America
Cell: 010-7103-8619
Email: kimsusi11@msu.edu; kimext@gmail.com

Hee-Kyung Lee
Graduate School of Education, English Education Major
Yonsei University
134 Sinehon-dong, Seodaemun-gu
Seoul 120-749, Korea
Phone: 02-2123-6265
Email: heelee@yonsei.ac.kr

Received in December 1, 2014
Reviewed in January 15, 2015
Revised version received in February 15, 2015