Talking Expert Raters and Rating Tasks

Lee, Ho*

〈ABSTRACT〉

The purpose of this study is to describe students’ ideas and attitudes about learner-directed rating, in which learners as raters participate in assessment activities. For this purpose, the researcher solicited eleven EFL students to participate in 40 minute interviews. Students responded that fairness and accuracy are the key factors for the expert raters. The way teachers respond to learners also deeply affects the perception of a expert rater and a test. After participants conducted multiple rating tasks, they demonstrated slightly positive attitudes toward their engagement in the rating tasks. However, they are concerned about the accuracy of their judgment in terms of linguistic components. In conclusion, the collaborative rating procedure, including learners’ participation in the rating process, should be carefully implemented within of the EFL context.

Key words: learner-directed rating, collaborative rating, L2 writing assessment

I. Introduction

Interest in measuring English writing and speaking has been raised recently in Korea. TEPS (Test of English Proficiency developed by Seoul National University), one of the prevailing English language tests developed in Korea, now includes TOP (Test of Oral Proficiency in English) and TWP (Test of Written Proficiency in English). In addition, Sookmyung University developed MATE (Multimedia Assisted Test of English), which is

* 중앙대학교 사범대학 영어교육과 전임강사
designated to measure speaking and writing skills. Second language researchers also have begun to publish their research papers to study the validity of such performance-based tests (Shin & Jang, 2002; Shin, 2001a; Shin, 2001b; Choi, 2001; Choi, 2000).

Administration of such performance-based tests comes with the demands for qualified raters. Since measurement of students’ performance is a time-consuming, elaborate, and sophisticated task, English testing institutions seek a number of qualified raters. They also need at least two qualified raters for each student’s speaking and writing sample as an effect to enhance test credibility and warranty. Moreover, they prefer English-native expert raters to non-native ones because of the general expectation that English-native speakers will show better rating performance than non-native ones. However, they are confronted with the very limited number of qualified native raters residing in Korea.

Coping with this problem, language-testing institutions conduct rater training for non-native English teachers or professionals. For instance, MATE requests applicants to complete an 8-hour workshop in which they are guided to be fully accustomed to MATE’s scale and scoring rubric. After the applicants successfully complete the workshop, they then take a test developed for measuring the rating capability. Upon its result, they are officially recognized as expert raters or, alternatively, are offered the training again. By using these steps, even a non-native speaker can be educated to be an expert rater.

In line with rater training for Korean raters, Shin (2001c) became aware of limitations within existing scoring guidelines as applied to English language tests for Korean examinees. He realized that familiar criteria such as ACTFL (American Council on the Teaching of Foreign Language), and procedures such as SOPI (Simulated Oral Proficiency Interview), do not fully provide detailed and comprehensive rating guidelines for Korean raters. Such ACTFL-SOPI guidelines, he suggested, are predominantly written for qualified native raters. Further, he insisted that scoring rubrics should be developed within the Korean learning and teaching context.

What about the test takers? Unfortunately, their voices have been ignored in test construction and implementation in Korea. Even though students’ responses or feedback collected from research settings partly affect the revision of an English test, opportunities for student participation in testing construction are sporadic, episodic, and limited. Rather, Korean EFL test takers as a whole have been considered at objects to be analyzed statistically.
Talking Expert Raters and Rating Tasks

As an alternative, the implementation of collaborative ratings can possibly relieve this problem. As collaborative ratings promote learner participation in test construction and development, students are able to provide explicit student voices about the testing setting directly to test developers. Language testers are also able to detect problematic assessment criteria from the perspective of test takers. Lee (2005) found that some Korean EFL students performing multiple rating tasks were willing to make corrective feedback on confusing descriptors of scoring guidelines at the time of interview. In addition, more than fifty percent of students responded that if requested, they would volunteer to rate essays as a complementary rater on a large-scale writing test. These findings suggest that rating data gathered from a large pool of student raters may make it possible for testers to be accessible to the enriched, comprehensive, and examinee-oriented information in test development.

Considering the benefits of the collaborative rating procedure, this paper is to situate our perspectives into the test takers’ eyes, supposing that the role of a student is transformed into that of a rater. The approach that the current study takes for addressing research questions is descriptive and learner-oriented. That is, it explores what test-takers have in mind when they encounter a rating task. Research questions that are followed include:

1. How do learners perceive the way expert raters assess the former’s performance?
2. What attitudes do learners holding regarding rating tasks after they perform collaborative rating?

II. Methodology

This section includes two sources: semi-structured interviews and email correspondence. These two types of data are collected from eleven students and three expert raters. The interview data are compiled according to loosely-categorized thematic schemes, so that interviewees’ views are described fully and with minimal intervention. Therefore, there are no analytic tables and figures. The data are presented according to two broad entries: pre-task level and post-task level.
Ten students participated for in the semi-structured interview. Interviewees were selected from those students who showed deep interests in rating tasks by leaving meaningful comments on the questionnaire. There were three interviewees from critics of the collaborative rating tasks, three interviewees from positive respondents, two interviewees from students stressing the difficulty of the rating tasks, and two interviewees who expressed their rating confidence.

Questions were developed for the semi-structured interviews. The format of all questions was open-ended, so students would be encouraged to make elaborated responses. They were thematically arranged, so students would not be confused. They were grouped by three subparts during collaborative rating: individual-independent rating process, negotiation process, and students’ attitudes/opinions toward the rating tasks. Questions might vary among interviewees, and some in-depth questions were designed only for certain students, in the case that he/she made interesting responses.

III. Results and Discussion

1. Pre-Task Level

A pre-task level analysis involves student perception of rating tasks before they conduct multiple rating tasks. It shows general beliefs, opinions, and feelings that students have historically had about rating. The pre-task level analysis includes only interviews with students.

Of 10 interviewees, Jung suggests a comprehensive definition of an expert rater. Here is a selected part of what he thinks a tester should be:

I think...a rater can be an instructor. In order to be a good rater, a person should know what is the ultimate learning goal of a target discipline to be assessed, what should be tested, … and how students have learned a target knowledge to be assessed. Based on such information, the tester may want to teach something. Students also want to learn something. The rater needs to have a standard that includes properly both standpoints, what he/she wants to teach, and what the learner wants to learn. The rater should have expertise in the specific domain. Moreover,
he/she should also be able to use the test results appropriately, not making a mechanical interpretation.

Looking closely at the translated transcription, Jung has high expectations for an expert rater. Jung stresses that a rating very often fails to indicate true differences among students in the English language test context. According to his opinion, a rater should at least endeavor to measure a student’s true performance with a variety of methods, and consider an examinee’s potentiality that a test is not able to reveal. In addition, he is also concerned about test use and test interpretation in high-stake and large-scale test contexts. He extends the role of a rater to monitoring the test use beyond test development and interpretation.

In fact, Jung does not make any distinction between rater and tester, even though I repeatedly asked him about his perceptions of an expert rater. Rather, he expects a rater to take charge of a variety of roles as a teacher and as a tester in order for the rater to understand students. He prefers raters who are able to observe students’ learning processes over the long term to those who measure student performance at a specific point. An ideal rater, as he opines, is an evaluator. Shin, another interviewee, also implies that rating is not an activity held only in classroom or test contexts. She mentions

Well, a rater… can be a parent or teacher in a specific situation. In general, the rater is like an expert who assesses students with reference to certain material.

One of the important requirements for the expert rater involves fairness. Some interviewees are very sensitive to the fairness issue. Therefore, they hope that a rater keeps distance from students, specifically those with high academic performance. They comment.

A rater should not be affected by personal bias for a student. A rating should be like a sword, which definitely cuts off the personal relationship with the students. (Min)

Rate fairly… A rater needs to strictly follow the standard that is not differentially applied for some of students. (Nam)

Most interviewees express that those who possess higher linguistic proficiency than students are entitled to be raters in a second language context. They stress that objective judgment comes from a rater’s expertise in the specific domain, English. Therefore, any
rater, they suggest, is required to know grammar and vocabulary very well. Here are their voices:

A rater should have a lot of experiences in the field that he/she rates. S/he reads a lot of essays and so predicts possible variables… The knowledge from a lot of experiences, that is important for the rater. (Yoo)

A rater has to be equipped with better linguistic proficiency than a test-taker. I lack the vocabulary and grammar proficiency with which to rate a peer. I can’t fully detect my mistakes with the limited capacity. Frankly speaking, such a lack of knowledge makes me not to pay attention to the peer’s grammatical and vocabulary. (Shin)

In many cases, interviewees report that they are not qualified as raters since they are not confident in English linguistic knowledge. Specifically, they ascribe the difference between themselves and the expert raters to the degree of proficiency in grammar and vocabulary. This feature of language use is a type of solid boundary to separate students from expert raters. Nonetheless, they do not fully preclude themselves from the position of raters. They still view organization as an area that they are able to assess. Some students further state:

Researcher: If you are requested to participate in an English essay test as an assistant rater, will you accept the offer?
Nam: I will hold back my decision if I am in that situation. If the content and vocabulary are difficult for me to read in an essay, then, it will be hard to rate the essay properly. I may not do well.
Researcher: If you rate an essay of peers with less proficiency?
Nam: In that case, it will be interesting.
Researcher: How about assessing a Korean essay?
Nam: It will be interesting, too.

Researcher: If possible, do you want to participate in a test as a rater?
Yoo: Well, as I indicated previously, I am not qualified as the rater. I will be sorry to an examinee. In a certain situation, the examinee can be a better writer than I in writing style and grammatical knowledge. How dare I assess him/her?
Researcher: You responded ‘strongly agree’ in the questionnaire.
Yoo: Assistant. Not as a main rater who give the determinant score. Though I don’t know English grammar, organization or something is a part that does not largely differ from a language to a language… I cannot assess grammar.
Talking Expert Raters and Rating Tasks

Researcher: If possible, do you want to participate in a test as a rater?
Na: I am not confident about rating. My scores did not match an expert’s during rating training.
Researcher: If you rate a Korean essay?
Na: Not an English essay... Yes. I will do it only in organization.

Another noticeable fact is that students’ awareness of their lack of proficiency causes them to conform to the authority of the rater, and the teacher. This phenomenon is partly associated with Confucian tradition, in that a student should respect a teacher to the highest degree of modesty. Here are the voices of students who show their compliance with the teacher rating:

Researcher: Did you experience that your performance was underestimated?
Seong: Yes. In that case, I talked to the classmates like, the score is wrong. That’s it. A teacher does not change the score, even if many students complain of their scores. Talk a word of complaint only one time, and that’s it. Some teachers admit our complaints, but take no action, just saying that’s too bad.
Researcher: Did you feel uncomfortable and angry?
Seong: No. I may have never felt angry and hostile. I tended to accept the academic standing in spite of the teacher’s problematic rating. I am not perfect. I am short of competence. So, I accept it because I know that I am not such a competent student.

Researcher: Did a teacher show the scoring criteria to you?
Na: No. He read a sort of reference, caught students’ errors, and gave scores. Scoring criteria was not open to students.
Researcher: Do you want to know the criteria?
Na: When I look around my classmates, good students got good scores. Most of all, I believed in the teacher’s rating ability, the teacher’s rating criteria. A teacher had more knowledge than I and peer classmates. I believed in her, certainly.

On the other hand, some interviewees show very negative perceptions of the teacher’s rating. In my individual interviews with Nam and Hyun, they are frustrated by the way in which some teachers treat them. The teachers not only keep the scoring criteria secret, but also blame the students for appealing the scores. As Bae (2004) suggested, the Confucian context that "articulating one’s opinion in front of the teacher is considered a challenge to authority."(p.346) may influence the teacher’s perception. According to Bae,
“in such an environment, students often learn to conform to the authority rather than raising questions or posing problems.” (p.346)

The teachers’ negative reaction affects and changes two interviewees’ perspectives on the rating:

Researcher: In the performance assessment, did you wonder how your performance was rated?
Hyun: Sometimes, I did… In the days of my high school, the scoring guideline was not public to students. I was only accessible to my holistic score. Sometimes I endeavored to write something elaborately. But I failed to meet the cut-off score. When I got B and B+, teacher said, ”You should resubmit a revised writing work.” At that time, I asked a teacher, ”Why did I get this score?” with some sense of disobedience that a guy usually feels in the adolescent… In the writing context where I should write my own perspective like an argumentative essay, in designing an essay structure, in designing the structure of introduction/body/conclusion, the essay with the specific point of view, you know, that a rater supports is likely to promise good scores. When I complained a teacher for the problem, what I got as a result of the complaint is a teacher’s negative reaction, something like, ”Why are only you so picky?” Since I met a dogmatic teacher, who pushed students to ”just do it because I do it,” I have been rather disinterested in how my performance is assessed. Because, in spite of my several requests, any teacher did not allow me to know how to score my performance, I gave up.

Researcher: Did you conform to the authority?
Hyun: I yielded to the authority rather than confirmed to it.

Researcher: Have you ever complained to a teacher about your underrated score?
Nam: Yes. But, (when I met a teacher,) the teacher hit me in the brow. Still, he/she did not correct the score. I hate the teacher. He/she did not have any specific criteria for rating. I did not tolerate against the teacher’s bias for the students with high academic records. So, I transferred to another school because I did not like the school.

Researcher: Did such an incident change your perception concerning rating itself?
Nam: I got a lesson that if a rater is not fair everything is in vain. I felt very frustrated.

In summary, students have their own frames to view an expert rater and ratings. Some are suspicious of the ways in which they are rated in real classroom settings. Sometimes such a suspicion remains unresolved in front of the teacher’s authority. Other students also show a high degree of conformity to the authority. In addition, a majority of students view linguistic proficiency as the pivotal requirement as a rater. The lack of
Talking Expert Raters and Rating Tasks

grammar and vocabulary knowledge, in their thinking, mostly contributes to creating a
distance between the students and an expert rater. Most of all, many or most students
possess a stereotyped image for an expert rater. They consider a rater as a teacher or a
professor who is able to understand their academic progress in the classroom context.

2. Post-Task Level

A post-task level analysis involves students’ overall evaluation about the rating tasks
that they complete for the research. It also includes the experts’ reactions regarding the
future use of the students’ rating tasks. The post-task level analysis shows two sources
of data, which are interviews and email communications.

Upon my request, three expert raters sent me emails about their perspectives of the
rating tasks. Sohn and Betty, as EFL raters, are experts who rated 104 student essays
for the current research. Sue rates a lot of English essays that students write for the
ESL Placement Test (EPT) at the University of Illinois in Urbana-Champaign, but did
not participate as an expert rater in rating the aforementioned 104 subjects. She was
instead requested to make comments about the collaborative rating procedure.

All of the expert raters show skeptical views about student participation in rating
tasks in large-scale test situations. Two of them think that rating tasks are somewhat
feasible in the classroom context under specified conditions. One expert, Sohn, denies the
effectiveness of the rating tasks. Here are the experts’ voices from the most negative to
the most positive response:

Students’ self/peer evaluation is an important part of learning. However, if its
objective is to rate their work, I do not believe it is an effective method, simply
because students lack the skills involved in rating. (Sohn)

I am skeptical that students could rate their own essays and peer essays without
prior training. If students are trained to do it, then it seems more possible. Based on
my teaching experience, I have found that most students don’t know how to evaluate
their own writing or the writing of peers without explicit training. On a large-scale
test, it would be impossible to know if students were using equivalent criteria in
their rating, nor could you know that they interpreted any grading guidelines in a
consistent way. (Sue)
Student assessment of their writing that uses the same criteria that the rater uses would be a good teaching tool. If used at the beginning and ending of the writing course, they could see their improvement. I like the idea. However, the criteria might have to be simplified for the student assessment. It would take time to teach at the beginning of the course. It is an interesting idea. I’m not sure how it would work for a large scale test. (Betty)

In contrast to discouraging responses by the experts, some students seem deeply motivated by the rating activities, though they think rating is a difficult task. Two students strongly express the interests in rating tasks:

Researcher: Do you think that you are able to be an expert rater?
Joo: Sure! I hope to be. I am not sure to be an expert rater in the future, but I want to become the rater. Definitely it is true.
Researcher: Why do you think you are not an expert rater at present?
Joo: What flashed upon me is grammar assessment. I am not totally competent in grammar and vocabulary.

Researcher: If possible, do you want to participate in the test as a rater?
Jung: I want to do. That will be wonderful experience. The rating task will be helpful for me. I should study very hard until I participate in the test as a rater. I need to practice writing a lot. I need to improve vocabulary, linguistic expression, and practice English conversation.

Furthermore, students from the interview witnessed the effects and consequences of the rating tasks they conducted. Although a majority of interviewees report that a single rating task does not greatly change their views on the rater and rating process, they have come to understand the role of an expert rater, as described in the following:

I wanted to have more opportunities for the rating. It was very good for you to ask us to rate essays. We have no further chances to rate essay any longer.
⋯ I have thought of a rater as a person who simply gives a score. Now, I think that the rater is an advisor who suggests diagnostic information. (Min)
Rating is a very difficult job. A rater rates an essay with reference to the very detailed sub-criteria of writings. I know any rater can rate it only after he/she completes rater training program. (Kang)

After all interviewees received the score report that included self-, peer-, and expert scores online, three of interviewees responded in emails. They all found that the experts rated their essays in a more severe pattern than expected. One student, Jung, promptly
adopted, the experts’ ratings. Two of interviewees, Seong and Min, expressed softened disappointment. Seong and Min, however, accepted the peer-rating as reasonable. Overall, all email respondents provided justification regarding their own ratings.

Here are students’ comments in their emails:

I was satisfied with the scores since both expert raters gave more than score 3 in the holistic component. I am okay about the discrepancy between the self-score and expert-score since I do not measure myself thoroughly. I should study hard. I became aware of the fact that I have potential as a good writer. I can feel it. I thought that although I am lack of grammatical knowledge, I am good at content and organization. From this report, I need to train more in content and organization. (Jung)

I felt somewhat shame since I found I overrated myself. I thought I did good job at content. But, the experts’ scores disappointed me a little bit. I speculated on the topic. Since I knew by myself how much time I spent for thinking content, I might give such an overrated score. (Seong)

I was a little bit disappointed at the score result given by two experts. Since I believed that my partner corrected the first self-scores accurately, the adjusted self-score would be exactly the same as the experts’. There might be a gap between the experts’ views and students’ views on the good writing. (Min)

The experts also provided comments on the score report that includes all types of ratings. The experts show neutral perspectives regarding the use of the score report, such as:

If I, as a student, received this type of feedback on a large-scale writing test, I would pay close attention to the expert rater and less attention to the feedback from the peer-rater. (Sue)

I will not care for students’ possible reactions to the score report. I do not think that the students raise any complaints after they compare self- and peer-ratings to my own ratings. (Betty)

Reporting students’ self-evaluation and peer-evaluation along with my rating results do not affect my ratings in any way. (Sohn)

Overall, the students and experts have differing perspectives on rating tasks. The students show more optimistic attitudes toward the use of the rating tasks in the EFL writing context. Alternatively, the experts demonstrate neutral or negative responses about
students’ participation in ratings. The expert raters suggest that rating tasks are possibly conducted in the classroom context.

VI. Conclusion

In summary, interviews at the pre-task level illustrate that the external/cultural environment significantly affects students’ perception of the expert raters or the rating tasks. While some interviewees admitted that the teachers and professionals are entitled to rate the quality of the learners’ performance, they also complained about the unfairness of the teacher ratings. Students at the post-task level observed that they should consider a variety of cognitive and social factors for more accurate judgment, which causes the rating tasks to be difficult. The comparisons of students’ responses to experts’ demonstrate that the two groups have differing expectations regarding the rating tasks. Compared to the experts, students have more positive views about the rating tasks.

This paper discusses how the collaborative rating procedure is applied to real-world contexts. Considering all interview analyses, I argue for the introduction and the administration of the collaborative rating procedure in the EFL Korean context. Students’ positive attitudes toward their participation in the rating tasks suggest that the collaborative ratings may be successfully administered in the classroom and/or in small-scale diagnostic test situations.

Collaborative ratings, however, should be carefully implemented in consideration of the EFL context and culture. In particular, collaborative ratings may conflict with the Korean testing culture in which a teacher’s evaluation is highly respected but student evaluation is not encouraged. Since test takers are not familiar with such collaborative tasks, they may feel uncomfortable about them. In addition, collaborative ratings undermine the tester’s authority in Korea. That is, when exposed to collaborative ratings, test takers may often question the validity, authority, and quality of a test, even with absurd excuses.

This paper has focused on students’ perspectives regarding rating tasks and expert raters. Even though it does not thoroughly depict learners’ views, it argues that we need
to actively accept learners’ voices in the test development process. This is a small but very important step to incorporate learner-centeredness in performance test contexts.
References


요약

평가전문가와 평가행위에 대한 영작문 학습자의 인식 조사: 인터뷰 분석을 통하여

이호

이 논문은 크게 두 가지 단계를 가지고 이루어져 있다. 첫째는, 영어 작문 학습자가 평가 전문가 및 평가 자체에 대하여 어떠한 인식을 갖고 있는지를 알아보는 사전 조사 단계가 있다. 두 번째로는, 학습들이 직접 영작문 채점 과정에 참여한 이후에 학습자의 평가참여에 대한 소감을 알아보는 사후 조사 단계가 있다. 이 두 단계의 고찰을 통하여, 학습자의 평가에 대한 인식을 알아보는 것이 본 논문의 목적이다. 본 논문을 위하여 모두 11명의 학생들이 40분간 진행하는 준 조직적 인터뷰에 참여하였으며, 그들의 인터뷰는 연구 목적을 위하여 녹음기로 녹음한 이후에 일일이 문자화 하였다. 학생들은 인터뷰를 통하여 교사의 교수행위와 같은 외부적인 요소와 문화적인 요소가 학습자의 평가와 평가자에 대한 인식에 대하여 중요한 영향을 끼쳤다고 대답하였다. 그들이 직접 영작문 채점 과정에 참여한 이후에는 평가 행위에 대하여 좀 더 긍정적인 시각을 갖게 되었음을 밝히었고, 영작문 채점 과정에 대하여 다소는 긍정적인 반응을 보였다. 반면, 평가 전문가들의 경우에는 학습자들의 채점과정에 참여에 대하여 회의적인 시각을 표출하였다. 본 연구를 통하여, 평가는 더 이상 학습자들과 유리된 개념이 아니며, 학습자들에게 능동적인 학습의 한 형태가 될 수 있음을 제시하였다. 다만, 학습자의 평가에 대한 참여는 현재의 한국적인 상황을 고려하여 신중하게 진행되어야 한다는 결론을 내렸다.

주제어: 영작평가, 협력평가, 학습자 위주 평가