# Reassessing "Success" in Peer Assessment

Zachary Robertson
*Yamaguchi University*

**Reference Data:**

Robertson, Z. (2025). Reassessing "success" in peer assessment. In B. Lacy, R. P. Lege, & M. Swanson (Eds.), *Moving JALT Into the Future: Opportunity, Diversity, and Excellence*. JALT. https://doi.org/10.37546/JALTPCP2024-45

Success in EFL peer assessment activities has traditionally been measured through adoption rates, with cases of feedback rejection viewed as either a failure of assessed students to incorporate advice or a lack of quality feedback. Researchers, however, have begun to question this premise and are now investigating the potential growth in evaluative expertise that such rejection could signify. This qualitative study of 54 university students enrolled in an EFL academic writing course in Japan examined how learners responded to feedback they received during peer assessment, focusing on the reasoning behind feedback rejection. Results indicate that rejection has less to do with disagreement with feedback rationale and more with students applying their evaluative expertise when deciding whether to use the feedback from their partners.

EFLにおけるピアアセスメント活動が成功であるか否かについては、これまで採用率を基準として測定され、フィードバックの拒否は、評価対象者が助言を取り入れることができなかったか、またはフィードバック自体の質の欠如と見なされることが一般的であった。しかし、研究者たちはこの前提に疑問を抱き始め、こうした拒否が評価スキルの潜在的な成長を示している可能性について調査を進めている。本研究では、日本の大学でEFLアカデミックライティングコースを履修する54名の学生を対象に、ピアアセスメントで受けたフィードバックに対する反応を質的に分析し、フィードバックを拒否する理由を検証した。その結果、拒否の理由として、フィードバックの根拠に対する不同意よりも、フィードバックを採用するかどうかを判断する際に、自身の評価者としての経験に基づいていることが明らかになった。

Peer assessment (PA), a popular assessment method in EFL and EFL contexts, is becoming a focal point for researchers as its relative strengths and drawbacks become better understood. Its proponents point to the cognitive and motivational benefits the activity fosters by having learners assume roles typically reserved for teachers (Grainger et al., 2018), as well as a social component requiring both assessors (those giving feedback) and assessees (those receiving feedback) to negotiate meaning and manage their interaction in a spoken and/or written communicative context (Yakame, 2005). Critics, on the other hand, have raised several concerns with the activity which include the quality/appropriateness of learner feedback, lack of cultural fit, and issues with interpersonal dynamics (Allen, 2015; Foster & Ohta, 2005). This has resulted in a pedagogical stalemate, leaving teachers to sift through the two sides of research to determine if the activity is appropriate for their classroom.

In response to this impasse, researchers are beginning to question the efficacy of metrics such as feedback adoption rate to judge the effectiveness of PA (Robertson, 2024). Leijen (2017), for instance, argues that rejection of feedback can be viewed as a sign of learner development when observing that "alteration, therefore, might not always lead to an improvement of the text, if students do not consider them carefully, but blindly implement them" (p. 48). Rather than a sign of poor or inadequate feedback, such rejection could actually indicate that learners are becoming more capable of judging quality in relation to activity objectives. It is therefore crucial to investigate this under-examined aspect of PA from a process-oriented perspective to better understand how assessor/assessee roles impact learner reactions to and perceptions of feedback throughout all phases of the activity.

## Theoretical Framework

The key to this investigation lies in differentiating the activity process (the way students learn throughout the activity) from activity products (i.e. learner feedback and produced works) and establishing a research apparatus capable of identifying formative growth in the learner. With a focus on formative development over summative outcomes, Carless's (2007) Learning Oriented Assessment framework has received attention from language educators by providing a theoretical foundation for reimagining

activity success (Wicking, 2022). The framework is built on three pillars that are used to determine if an assessment model is functioning in a formative manner: 1) fostering sound learning practices, 2) promoting student engagement with feedback, and 3) the development of evaluative expertise (Carless, 2015). Of these, the pillar of evaluative expertise is of particular interest because it converges directly with PA, which requires students to demonstrate competency in all aspects of evaluation. This includes synthesizing the goals of the initial task, learning targets associated with said task, explicit criterion for judging quality, and implicit knowledge of relevant exemplars (both positive and negative) to form a comprehensive judgment of the work they are evaluating (Sadler, 1989). From a process standpoint, learners essentially are comparing what they are judging to pre-constructed quality exemplars and, through the reconciliation of these discrepancies, solidifying their understanding of relevant quality criteria (Grainger et al., 2018).

After applying this framework to PA from an operational standpoint, several alternatives to reinterpreting feedback rejection emerge. The first and most immediate example is learner dissatisfaction with the feedback they have received from their peers due to perceived flaws (the feedback contains errors, disagreeing with the assessor's rationale for the feedback, etc.) in the feedback itself. If feedback quality is the sole determining factor for evaluating activity success, the byproduct of the activity (the feedback) and downstream artifacts (final activity product) become the only windows into learner development. This, however, overlooks the evaluative expertise that learners must acquire to even make the claim that the feedback they received is inadequate. In other words, before they can reject feedback on the basis of quality, they must first internalize what "quality" means in respect to the task in which they are engaged. Were such a rejection to be based on a rational decision-making process, it should be viewed as a positive outcome rather than a failure of the activity.

In addition, there is another possibility that feedback rejection has less to do with the quality of the feedback and more to do with learners applying their evaluative expertise when performing their roles as assessors. Cognitive models of skill acquisition conceptualize complex skills like evaluative expertise as a bundle of modularized micro-skills that, upon sufficient mastery, can be applied to novel situations (Mascolo, 2020). In the context of PA, Robertson (2024) describes a hypothetical scenario in which Learner A receives feedback from Learner B to rectify a problem with their work. Rather than acting on Learner B's suggestion, however, Learner A remembers a similar example from work they saw while assessing the work of Learner C which better addresses their problem and opts to use that instead. Normally, this would be classified as feedback rejection and understood in negative terms. However, this does not account for the transferability of evaluative expertise to all aspects of the activity, which allows learners to tap into their experience as assessors when determining the best option to improve their work.

Bearing this in mind, the following avenues of investigation into learner perceptions and responses to feedback will be pursued in this study:

1. Does learner approval of partner feedback decrease as learners gain more evaluative expertise?
2. Is there a connection between evaluative expertise and what assessors focus on when constructing feedback?
3. Do learners reject feedback based on disagreements with the rationale of their assessor?
4. To what extent does the assessor role influence feedback rejection?

This exploratory study will investigate these topics via a qualitative approach to be outlined in the following section that, if successful, can be used as the foundation for a more rigorous mixed-method investigation.

## Method

### Data Sample

The study sample included 54 3rd-year university students enrolled in an elective academic writing course at a national university in Japan. The learners, all engineering majors, were a mix of Japanese ($n$=38) and international students ($n$=16) from China, South Korea, and Malaysia. In terms of language proficiency, the class possessed TOEIC L&R scores between 450~950, which translates roughly to B1~C1 on the CEFR scale (ETS Global, 2021) and represents a significant range in learner ability. Scores were collected and made available to the study author by the institution, which set a minimum score requirement of 450 to take the class. At the start of the course, all students signed an informed consent form after receiving an explanation in both English and Japanese of the purposes and goals of the study, which was authorized by the university's institutional review board.

### Procedure

The study was conducted over a 15-week period in 2024 in which students completed four 300-word essays on different topics (personal statement, body language, project

analysis, and concept explanation), with a PA session between the submission of the first and second drafts. Each assignment cycle took place over three weeks and consisted of a first draft submission, a PA phase in which feedback was graded by the instructor on a 0-5 scale, and finally a second draft submission that was screened for potential AI abuse by the instructor using specialized detection software before being assessed. At the end of the semester, students then selected their three best works for submission in a final portfolio that represented 40% of their overall course grade. In order to incentivize student participation throughout the entire activity cycle, drafts (10%) and PA feedback (20%) were included as part of the overall course grade.

PA sessions were conducted online using the Moodle 3 workshop platform, allowing the process to be both randomized and anonymous. For the first activity cycle, students were allowed to freely assess the essay of a single peer with no requirements other than that their feedback had to be written in English and contain a minimum of 100 words. Upon completion of the first activity cycle, learners then completed a 45-minute feedback instruction session where they learned Min's (2005) "Clarify-Identify-Explain-Offer" method to generate feedback, which was then used as the basis for grading in the remaining three activity cycles. The reason for conducting the instruction session after the first activity session was to test for potential discrepancies in evaluative expertise between the first and following weeks of the study.

### Instrument

Following each PA session, learners were then required to complete an online (Moodle 3) survey related to the feedback they received before submitting their second draft. The survey consisted of four multiple-choice questions written in both Japanese and English, with each version verified by a native speaker to ensure abstract concepts like "global" and "local" were properly communicated to the students. To encourage learners to be as honest as possible, only the study author had access to the survey results, and learners were assured that their responses would not affect their grade or the grade of their partner.

Survey questions were designed to address the four areas of investigation introduced in the previous section and are presented in their English version below:

1. On a scale from 0-5, how helpful was the feedback you received from your partner?
2. Did the feedback you received focus primarily on GLOBAL (ideas and organization) or LOCAL (grammar, spelling, punctuation, etc.) aspects of your essay?
3. What do you think about the rationale (reasons) your partner provided for their feedback?
4. How did you respond to the suggestion(s) your partner offered in their feedback?

At the conclusion of the study, responses were then aggregated and organized by learner TOEIC score into three cohorts of roughly equal size: 450~550 ($n$=19), 555~700 ($n$=17), and 705~950 ($n$=18). This, as with the assessor training, was done to screen for potential effects related to discrepancies in evaluative expertise within the population.

## Results and Discussion

The following section discusses the results for learner feedback grades (the grade given by the instructor) and the four survey items for each TOEIC cohort.

### Learner Feedback Grades

**Table 1**

*Learner Feedback Grades (0~5)*

| TOEIC Cohort | TOEIC Average | Cycle 1 Grade | Cycle 2 Grade | Cycle 3 Grade | Cycle 4 Grade | Overall |
|---|---|---|---|---|---|---|
| **450~550** | 510 | 4.67 | 4.25 | 4.33 | 4.60 | 4.47 |
| **555~700** | 610 | 4.64 | 4.50 | 4.40 | 4.75 | 4.52 |
| **705~950** | 845 | 4.82 | 4.77 | 4.88 | 4.77 | 4.81 |
| **All (54)** | 660 | 4.71 | 4.51 | 4.58 | 4.71 | 4.63 |
| **Pearson (*r*)** | | .15 | .34 | -0.05 | -0.05 | .22 |
| **p-value** | | .27 | .01 | .02 | .73 | .11 |

Table 1 shows a small positive relationship between TOEIC score and learner feedback grades, which is understandable given that language competence is a primary component of evaluative expertise. It should be noted, however, that scoring only considered whether students fulfilled the basic requirements for the feedback and not the quality of the feedback itself. This was done to remove as much subjectivity from the grading process as possible and resulted in overall high scores. There was a slight drop in scores in the second cycle, which was the first time the students were required to use the four-step assessment method they learned during feedback training. Scores then increased

over the remainder of the semester as students became more familiar with the method, lending credence to the hypothesis that evaluative expertise can be gradually acquired with intentional repetition. It also provides evidence that learner evaluative expertise was improving throughout the semester, which can be used to interpret the four survey item responses.

## Assessee Perception of Feedback (Item 1)

### Table 2

*Cohort Feedback Perceptions (1~5)*

| TOEIC Cohort | Cycle 1 Average | Cycle 2 Average | Cycle 3 Average | Cycle 4 Average | Overall Average |
|---|---|---|---|---|---|
| **450~550** | 4.44 | 4.44 | 3.63 | 4.39 | 4.36 |
| **555~700** | 4.62 | 4.33 | 4.19 | 4.57 | 4.31 |
| **705~950** | 3.87 | 4.00 | 3.70 | 4.25 | 3.78 |
| **All (54)** | 4.31 | 4.26 | 3.00 | 4.40 | 4.42 |
| **Pearson (*r*)** | -0.13 | -0.13 | -0.29 | .07 | |
| **p-value** | .34 | .36 | .037 | 0.59 | |

### *Question 1: On a scale from 0 (least helpful) to 5 (most helpful), how helpful was the feedback you received from your partner?*

The results from Item 1 suggest a moderately negative correlation between TOEIC level and perception of feedback utility, supporting the hypothesis that more advanced learners tend to be more critical of the feedback they receive. Although the limited sample size prevents drawing definitive conclusions of statistical significance for most individual cycles, Cycle 3 is of particular interest in that there is a clear dip in learner satisfaction across all TOEIC cohorts. This could suggest that learners are becoming more discerning as they gain experience with the evaluation process, or it may simply be a byproduct of that particular assignment. The fact that the scores picked up again in the fourth cycle suggest the latter, but a longer investigation would be necessary to confirm or reject that claim. If anything, this shows that learner opinion of peer feedback can fluctuate significantly from activity to activity, with the most extreme swings coming from the higher ability students.

## Feedback Target and Assessee Perception (Item 2)

### Table 3

*Feedback Target Totals (%) and Correlation to Feedback Perception*

| Response | Cycle 1 | Cycle 2 | Cycle 3 | Cycle 4 | Average |
|---|---|---|---|---|---|
| **Global (1)** | 60.00 | 60.98 | 50.00 | 56.41 | 56.97 |
| **50 /50 (2)** | 28.89 | 19.51 | 27.50 | 23.08 | 24.85 |
| **Local (3)** | 11.11 | 19.51 | 22.50 | 20.51 | 18.18 |
| **Pearson (*r*)** | -0.68 | -0.74 | -0.51 | -0.70 | |
| **p-value** | < .01 | < .01 | .01 | < .01 | |

### *Question 2: Did the feedback you received focus primarily on GLOBAL or LOCAL aspects of your essay?*

While not the primary focus of this study, the above results can help explain the sudden drop in learner perceptions during Cycle 3 of Item 1 by providing additional context to the feedback learners received from their peers. Given that the options for this question were numbered *Global* as "1," *50-50* as "2," and *Local* as "3," learner perception of feedback can be seen roughly correlating with the amount of global feedback they received (the negative Pearson correlation implies learners who answered higher in Item 1 chose Options 1 and 2 more than Option 3). This suggests learners were seeking analysis from their peers that extended beyond the superficial level of local feedback, which they subsequently assigned a lower rating.

What is key here, however, is what is driving this response: learner expectations of the assessment or the actual quality of the feedback itself? The next survey item will address this question by examining assessee perception of the rationale provided by their partner.

## Perception of Assessor Rationale (Item 3)

Table 4

*Assessee Perception of Assessor Rationale (%)*

| Response | Cycle 1 | Cycle 2 | Cycle 3 | Cycle 4 | Overall |
|---|---|---|---|---|---|
| **Agree** | 91.11 | 90.24 | 97.50 | 97.50 | 93.98 |
| **Disagree** | 6.67 | 2.44 | 0 | 0 | 2.41 |
| **Did not understand** | 2.22 | 2.44 | 2.50 | 0 | 1.81 |
| **Not provided** | 0 | 4.88 | 0 | 2.50 | 1.81 |

### *Question 3: What do you think about the rationale (reasons) your partner provided for their feedback?*

The results of this question item indicate that not only did students agree with the reasoning provided by their assessors but also that the agreement rate increased over time, indicating students were becoming more adept at using the four-step approach taught during assessor training. This is an important point to establish, as it rules out a potential reason why students might reject or look unfavorably upon the feedback from their partner. It may not provide answers as to the quality of the feedback, but it does show that assessees generally agreed with their partner about what needed to be changed and offers additional evidence of learners preferring deeper global analysis over local analysis. Most importantly, it can be used in interpreting responses to the final question item, which asks students how they utilized their partner's feedback.

## Assessee Responses to Feedback (Item 4)

Item four offered nine choices representing the possible actions learners took in response to the feedback they received, as well as their reasons for doing so. These choices (the results of which are listed in Table 5) are numbered as follows:

1. I used the suggestion(s) they offered with few or no changes.
2. I used the suggestion(s) they offered after adjusting it to fit my content.
3. I did not use the suggestion(s) they offered, but the suggestion(s) gave me an idea that I used instead.
4. I did not use the suggestion(s) they offered because I found a better solution while I was reviewing another student's work.
5. I did not use the suggestion(s) they offered because I couldn't think of a way to apply it to my essay.
6. I did not use the suggestion(s) they offered because they were either inappropriate or contained too many errors.
7. I did not use the suggestions they offered because I felt the tone of the feedback was too negative/impolite.
8. I did not use the suggestions they offered because I did not think any changes were necessary.
9. My partner did not provide any suggestions to improve my essay.

Table 5

*Assessee Response to Feedback (%)*

| Choice # | Cycle 1 | Cycle 2 | Cycle 3 | Cycle 4 | Overall |
|---|---|---|---|---|---|
| **1** | 2.22 | 2.44 | 10 | 12.82 | 6.67 |
| **2** | 80 | 82.93 | 77.50 | 61.54 | 75.76 |
| **3** | 11.11 | 4.89 | 7.50 | 15.39 | 9.70 |
| **4** | 0 | 0 | 2.50 | 2.56 | 1.21 |
| **5** | 2.22 | 2.44 | 0 | 0 | 2.42 |
| **6** | 0 | 2.44 | 0 | 0 | 0.61 |
| **7** | 0 | 0 | 0 | 0 | 0 |
| **8** | 0 | 4.88 | 2.50 | 0 | 1.81 |
| **9** | 4.44 | 0 | 0 | 2.56 | 1.81 |

### *Question 4: How did you respond to the suggestion(s) your partner offered in their feedback?*

Choices 3~8 deal with situations when assesses did not adopt the suggestions from their peers, and as a group they fall in line with previous research reporting rejection rates of 10%~20% (Saito, 2008). All but Choice 7 indicates a discernment of quality/appropriacy on the part of the assessee, with Choices 3 and 4 in particular revealing

a direct connection between the learner's role as assessor and their response as an assessee. There is a noticeable jump (~10%) in these two choices during the final cycle, but it is impossible to know if this is the beginning of a trend of assessees exercising their evaluative expertise more actively or simply a byproduct of that particular writing assignment. A longer study with more PA cycles would be necessary to confirm if these two reasons increased with learner repetition but given the low rates for Choices 6~8, it can be argued that assessee rejection should not be immediately interpreted as evidence of poor feedback.

## Study Limitations

Although the collected data shone light on several aspects of feedback rejection, several factors prevent broad generalization of the data. The first is the limited sample size and length of the study, which allowed for only four PA cycles to draw from. Another is the lack of quantitative data to validate the results of the survey, which drew exclusively from the subjective experience of learners who sometimes give idiosyncratic or contradictory responses to question items. This lack of data applies to both the feedback learners received as well as the first and second drafts of the assessees. To be certain, a full qualitative analysis of all drafts and feedback would require a significant amount of time to be conducted properly, but emerging AI technology (see ChatGPT, for example), though still unproven as a research tool, could offer a possibility to reduce the burden of the researcher in the future.

The final concern has to do with the timing of the assessor training, which was conducted between the first and second PA cycles. This was done to test if any differences existed after conducting the peer assessment training, but on reflection it may have simply introduced another complicating variable without significant differences in learner responses to show for it. Had the training been conducted before the first cycle, it may have been possible to diagnose trends related to Item 4 of the survey with a higher degree of certainty.

## Conclusion

Returning the initial research questions of this study, there is evidence to suggest that learner approval of feedback depends more on the content of the assignment than evaluative expertise (Question 1) and that feedback rejection is generally not based on the rationale of the assessor (Question 3). The data did not, however, reveal a clear connection between evaluative expertise and what assessors choose to focus on when constructing feedback (Question 2), which could be further investigated by adding open-ended questions covering the assessor experience to the questionnaire. This could also provide additional clarity to the extent the assessor role influenced feedback rejection (Question 4). A follow-up study which incorporates the following changes in response to the caveats of the previous section could potentially address these concerns:

1. Leverage LLMs to analyze both pre- and post-activity learner drafts for more robust quantitative analysis.
2. Include a free answer section in Item 4 of the survey to uncover other reasons for feedback rejection.
3. Include a question item to address the assessor experience, such as confidence in the quality of their own feedback and the target of their feedback.
4. Conduct the assessor training before the first cycle to reduce the amount of complicating variables in the study.

## Bio Data

**Zachary Robertson** is an associate professor at the Yamaguchi University School of Engineering, where he teaches technical communication, academic writing, and engineering presentation courses at both undergraduate and graduate levels. His research interests include peer assessment, CLIL, and AI supported education. <zachary@yamaguchi-u.ac.jp>

## References

Allen, D. (2015). Personal and procedural factors in peer feedback: A survey study. *Komaba Journal of English Education*, *6*, 47–65. https://repository.dl.itc.u-tokyo.ac.jp/record/2000962/files/KJEE06_047-065.pdf

Carless, D. (2007). Learning-oriented assessment: conceptual bases and practical implications. *Innovations in Education and Teaching International*, *44*(1), 57–66. https://doi.org/10.1080/14703290601081332

Carless, D. (2015). Exploring learning-oriented assessment processes. *Higher Education*, *69*(6), 963-976. https://doi.org/10.1007/s10734-014-9816-z

ETS Global (2021). TOEIC® Listening and reading test scores and the CEFR levels. ETS Global. Retrieved December 11, 2024, from https://etswebsiteprod.cdn.prismic.io/etswebsiteprod/515ff84d-fdd0-4e7f-8d79-e19b4d19e9db_TOEIC+Listening+and+Reading+Mapping+Table+CEFR-092021.pdf

Foster, P., & Ohta, A. (2005). Negotiation for meaning and peer assistance in second language classrooms. *Applied Linguistics*, *26*(3), 402–430. https://doi.org/10.1093/applin/ami014

Grainger, P., Heck, D., & Carey, M. (2018). Are assessment exemplars perceived to support self-regulated learning in teacher education? *Frontiers in Education*, *60*(3), 1–9. https://doi.org/10.1080/02602938.2012.674485

Leijen, D. (2017). A novel approach to examine the impact of web-based peer review on the revisions of L2 writers. *Computers and Composition*, *43*, 35–54. https://doi.org/10.1016/j.compcom.2016.11.005

Mascolo, M. F. (2020). Dynamic skill theory: An integrative model of psychological development. In M. F. Mascolo & T. R. Bidell (Eds.), *Handbook of integrative developmental science: Essays in honor of Kurt W. Fischer* (pp. 91–135). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781003018599-4

Min, H. (2005). Training students to become successful peer reviewers. *System*, *33*(2), 293–308. https://doi.org/10.1016/j.system.2004.11.003

Robertson, Z. (2024).「結果より課程:ピアレビューにおいての評価専門知識を再評価する」[Process over product: Revaluating the role of evaluative expertise in peer assessment]. 中国地区英語教育学会詩 [*Casele Journal*], *54*, 39-52.

Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Language Testing*, *25*(4), 553–581. https://doi.org/10.1177/0265532208094276

Wicking, P. (2022). Learning-oriented assessment as a theoretical framework for exploring teachers' assessment beliefs and practices. *JALT Journal*, *44*(1), 57–80. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/jj44.1-art3.pdf

Yakame, H. (2005). The role of peer feedback in the EFL writing classroom. *Annual Review of English Language Education in Japan*, *16*, 101–110. https://doi.org/10.20581/arele.16.0_101