Language Research Bulletin, 34, ICU, Tokyo

Effect of Peer Feedback on Self-assessment in Writing

John Peloghitis
English for Liberal Arts
International Christian University

Abstract

Self-assessment and peer feedback have been widely recognized as valuable pedagogical tools to promote autonomy and motivation in second language learners. However, their effectiveness is undermined if they are poorly implemented and fail to foster the skills students need to evaluate their own work. This paper examines the impact that peer feedback has on helping students to assess their writing performance. Self-ratings and ratings by independent raters on five areas of writing were collected from two essays and examined in two groups of students over a year-long writing course: twenty students who assessed themselves after peer revision and twenty-three students who assessed themselves after self-revision. T-tests were conducted to determine if significant differences exist in the level of agreement between students and raters and if these differences decrease over time due to the presence of peer feedback. The data suggests that the group engaging in peer feedback made greater progress in overcoming significant gaps in agreement, particularly in overall essay score and grammar. The results indicate that more attention is needed to create opportunities to integrate peer assessment exercises to complement self-assessment as well as other types of alternative assessment in process-oriented writing programs.

A growing body of research in mainstream education and second language learning supports claims that self-assessment is an effective practice for improving student performance in writing (Khodadady & Khodabakhshzade, 2012; O'Malley & Valdez Pierce, 1996) and developing learner autonomy (Blanche, 1988; Honsa, 2013; Sambell, McDowell, & Sambell, 2006). The process of self-assessment has been defined as the involvement of learners in identifying standards or criteria to apply to their work and making judgments about the extent to which they have met these standards or criteria (Brew, 1999). Kavaliauskiene (2004) argues that this process gives learners opportunities to think about their own progress and find ways to change, adapt, or improve it. Black and Wiliam (1998) add that self-assessment in which "the desired goal, evidence about the present position, and some understanding of a way to close the gap between the two" (p. 7) is a critical process for future learning. Proponents of self-assessment indicate that learners can develop the skills and autonomy needed to evaluate their work when teachers are no longer present to support them.

However, despite the widespread belief among English as a Second Language (ESL)/ English as a Foreign Language (EFL) practitioners and researchers that self-assessment benefits learning, doubts persist that students are often ill-equipped or lack the ability and commitment to assess their writing in a valid manner (Ross, 2006). The validity of self-assessment, which refers to the level of agreement of scores from student self-assessments and scores from teachers or

raters (Magin & Helmore, 2001; Topping, 2003), has been addressed in a number of studies but the findings are mixed, and thereby inconclusive. Part of the problem stems from the complex and sometimes overwhelming process of learning to write in a second or foreign language, which leaves learners with vague notions about what and how they are to assess. More research is needed to explore which areas of writing (i.e., organization, grammar, and coherence) are difficult for learners to evaluate and identify systematic approaches that enhance the process. To help learners make valid self-assessments, researchers have emphasized the importance of training and repeated exposure (Nunan, 1996; Orsmond, Merry, & Reiling, 2000; Min, 2006). Others have claimed that peer feedback can be a valuable pedagogical tool in preparing learners since it creates opportunities for interaction and increases objectivity in their evaluation (Black, Harrison, Lee, Marshall, & Wiliam, 2003). By sharing information and analyzing texts from their peers, learners can become more aware of the assessment process and the criteria in which they will be judged. Before proceeding further, it is salient to distinguish the terms 'peer feedback' and 'peer assessment.' Peer feedback is a communication process in which learners generate dialogue and make comments related to performance and standards. However, peer assessment is the evaluation of work or performance using a set of relevant criteria (Falchikov, 1995). In other words, peer feedback is the learning component of peer assessment.

Factors Impacting the Validity of Self-assessment

Studies examining the effectiveness of self-assessment in establishing agreement between learners and teachers have provided mixed results. AlFallay (2004) and Cheng & Warren (2005) investigated self-assessment and reported a high level of agreement between the self-assessments by students and the assessments by independent raters or teachers. Williams (1992) also reports close agreement between self-ratings and teacher ratings when students had a reference available to help them rate themselves as did Dochy, Segers, & Sluijsmans (1999) when they examined students grading their own essays. Stefani (1994) analyzed the correlations between self- and tutor-assessment and found that the students' self-assessments closely matched the tutor's marks (r-value = 0.93). Oldfield and Macalpine (1995) and Sullivan and Hall (1997) have also reported high correlations between teacher ratings and ratings from self-assessments. These results suggest that students do possess the ability to make valid assessments of their work. Boud and Falchikov's (1989) meta-analysis of earlier research supports the idea that students can make valid self-assessments; however, they raised concerns over methodologies in many of the 48 studies they examined. In fact, some research studies have indicated low agreement between students and teachers (Jafarpur, 1991; Hughes & Large, 1993; Mowl & Pain, 1995; Orsmond, Merry, & Reiling, 1997).

The Role of Peer Feedback

With the adoption of more learner-centered approaches in language teaching, researchers have examined forms of peer feedback to investigate whether they help learners produce self-assessments that are more compatible with teacher scores. Patri (2002) examined the effectiveness of peer feedback on peer and self-assessment on oral presentations by Chinese students. In the study, Patri compared the level of agreement in self-assessment and peer-

assessment scores between a group of students using peer feedback and a group of students who did not use peer feedback. For both groups, t-tests and Pearson correlations revealed greater validity in their peer assessment scores ($r = 0.85$) than in their self-assessment scores ($r = 0.46$). Based on the findings, Patri suggests that peer feedback helped students to gain a clearer understanding of the assessment criteria resulting in more valid assessments.

In a similar study, Saito and Fujita (2004) examined scores on essay quality in Japanese undergraduate students. They examined the relationship between rater scores and self-assessment scores as well as rater scores and peer assessment scores. The data supports Patri's findings (2002) that students' scores on peer assessment, not self-assessment, are strongly correlated with scores from teachers or raters. The researchers contend that psychological and cultural factors such as self-confidence and modesty create problems for students to assess themselves accurately. With the exception of these two studies, there remains a lack of robust empirical evidence linking peer feedback with valid assessment. Despite the ambiguity, however, students still perceive peer feedback to be a valuable practice.

An additional study by To and Panadero (2019) explored the effects of peer assessment on the self-assessment process and the factors limiting the effectiveness of peer feedback in 11 first-year undergraduates. The researchers used a qualitative approach by examining students' journals, follow-up interviews, observations of in-class formative peer assessment activities, and teacher interviews. They found that peer feedback could aid the self-assessment process by enriching student understanding of quality, refining subjective judgments, and deepening self-reflection. The study also concluded that teachers must scaffold peer feedback carefully to reduce tensions in feedback communication and the lack of readiness for peer learning.

Peer feedback is often implemented in conjunction with some form of scaffolding, for example, rubrics and checklists. Rubrics, in particular, lead to more valid scores because students are less likely to overestimate their peer's work (Panadero, Romero & Strijbos, 2013). Even though scaffolding can improve both the quality of peer assessment and increase the amount of feedback assessors provide, it has not always lead to more accurate self-assessments (Peters, Körndle, & Narciss, 2018).

Although there is a wealth of literature on self-assessment, few studies to date have investigated changes in the validity of self-assessment over time. While the studies by Patri (2002), Saito and Fujita (2004), and To and Panadero (2019) examined the roles of peer feedback, these studies did not explore its effect on improving self-assessment, which is a fundamental step toward learner autonomy. Thus the present study seeks to determine whether peer feedback has an impact on the level of agreement between student self-assessment scores and rater scores and whether students can make more valid ratings after repeatedly engaging in peer feedback activities. A second and equally important impetus for this study is to observe if students have areas of their writing that are more difficult to assess than others and whether peer feedback experiences help them assess their own writing more effectively in these areas. Results can provide information for writing practitioners to target areas of assessment where significant gaps exist in the level of agreement between teachers and students. The study aims to shed light on the following research questions:

1. Do significant gaps exist between the student self-assessments and the assessment by teachers or raters? If so, in what areas of writing are these significant gaps observed?

2. Does peer feedback enable students to make self-assessments about their essays that are more comparable to the assessments by teachers or raters? If so, what areas of writing become more compatible over time?

Methodology

Participants

The participants in this study were 43 first-year Japanese students enrolled in a two-semester compulsory writing course. All of the participants were English communication majors attending a private university in Japan. The participants were classified as 'lower intermediate' based on a diagnostic exam administered by the university a week before the start of classes. It should be noted that a writing component was not included in the exam. Twenty-five of the participants were female, and 18 were male, and all reported little or no experience in process-oriented expository writing.

The instructional context

The primary goal of the writing course was to familiarize students with a process-approach and basic academic writing with particular attention on developing cohesive paragraphs, and organizing ideas into clear, logical compositions. To address the course objectives, students learned organizational and rhetorical structures commonly produced in academic environments, and submitted and revised multiple drafts of writing.

The participants were from two separate writing classes taught by the same instructor. Students in both classes met once a week for two semesters. Each class was 90-minutes, and students met 14 times each semester. Although students were required to hand in two assignments each semester (four in total), only the first and last assignments were used to make observations related to the research questions. For each assignment, each group was required to submit three drafts. The first drafts were selected to measure the compatibility between the rater assessments and the student assessments over the duration of the course. They were submitted in week seven in the first semester and week twelve in the second semester. Each group was required to submit three drafts for each assignment. The two essays were both argumentative essays that required students to clarify their position on a topic and present two reasons with supporting details. Students had to choose one of several topics provided by the instructor, and models were provided for guidance. All students were encouraged to write over 300 words for each draft.

Research design and procedure

A control group and an experimental group were established to evaluate the effect of peer feedback on self-assessment. The control group consisted of 23 students that edited their first drafts using checklists followed by self-assessment. The experimental group included 20 students who conducted peer feedback on their essay drafts, followed by self-assessment.

A number of suggestions made in prior research studies were considered to provide effective training and minimize the effect of intervening variables that may cause discrepancies in the level of agreement. See the Appendix for details regarding the training sessions and feedback process.

First drafts were returned with an attached handout that contained positive and

constructive comments as well as a preliminary score for each of the five areas of writing. Many of the comments targeted specific but global problems in the essays and included suggestions on improvement. The same procedure was followed for the second draft, but more emphasis was given to discrete issues. Raters evaluated the students' essays using a scale from 1 (low score) to 5 (high score) in five areas of writing; main ideas, organization/coherence, supporting ideas, grammar, and vocabulary. It is important to note that although the raters did score the essays, the instructor provided the comments and some indirect error correction, and the scoring and commenting on the essays were done independently. At the time of the study, the instructor had been teaching writing in various tertiary contexts in Japan for eight years. Two native-speaking English language instructors at the same university rated the essays to attain more objectivity. Both raters had more than six years' experience in writing instruction and were teaching the same course at the time of the study. Written consent was obtained from each participant at the outset of the study. Student numbers were used during the rating of the essays to ensure anonymity.

Analysis

The validity of self-assessment on each category was determined as the level of compatibility between the scores the raters assigned on each category and the scores the students assigned on their own essays. For the first research question, independent sample t-tests were calculated to report on any significant differences found between each group of students and the raters for the first essay. The second research question was investigated by observing the scores given in the five areas of writing on the first drafts of the first and last argumentative essays. These calculations were used to establish pre- and post-treatment measurements for the control group and the experimental group. Improvement in validity was measured by examining the mean differences between the raters' scores and students' scores, and if significant differences on the level of agreement in the first essay became non-significant on the final essay. The alpha for achieving statistical significance was set at .05. Additionally, effect sizes using Cohen's *d* were calculated on the t-tests to evaluate the stability and strength of significance.

Because two raters were used throughout the study, interrater reliability was measured. Reliability measures were first established using a Pearson product moment correlation coefficient on a random sample of essays before any evaluation or marking was performed on the essays. The overall computed Pearson correlation coefficient was strongly correlated for the first essay ($r = .731$, $p < .01$) and the last essay ($r = .779$, $p < .01$), which indicates that a strong relationship was found between scores assigned by the raters on both essays.

Results

The data included rater assessments and self-assessments from both groups obtained in week 7 in the first semester for the first essay and week 10 in the second semester for the fourth essay. The data in Table 1 is relevant to answer the first research question, which examines if significant differences exist between the self-assessments and the rater assessments. The data indicates that students in both groups scored themselves higher than the raters on every category on the first essay, which accounts for the gap in the overall difference (3.55 for the control group and 2.60 for the experimental group). Higher student assessments were also observed in the last essay, with the exception of organization, which was underestimated by both groups. The data

also reveals that the students from the experimental group received slightly higher scores from the raters, and they had a higher level of agreement in their self-assessments than students from the control group.

For the second part of the first research question, which investigated the areas of writing where the largest discrepancies exist between student scores and rater scores in the first essay, the results showed some similarities. Descriptive statistics in Table 1 show that the area with the lowest score for the control group was main idea (.91) followed by support (.83). Likewise, the areas with the largest discrepancies for the experimental group was support (.81) and main idea (.61). Each area of writing for the control group was, in fact, greater than the experimental group.

Table 1
Descriptive Statistics for the Control Group and Experimental Group on Both Essays

		Control Group (*n*=23)			Experimental Group (*n*=20)		
	Area of Writing	Mean Self-score (SD)	Mean Rater Score (SD)	Mean Diff.	Mean Self-score (SD)	Mean Rater Score (SD)	Mean Diff.
Essay 1 - First Draft	Main Idea	3.54 (.90)	2.63 (.71)	.91	3.38 (1.07)	2.77 (.73)	.61
	Organization	3.46 (.80)	2.74 (.56)	.72	3.3 (.85)	2.87 (.63)	.43
	Support	3.37 (.66)	2.54 (.60)	.83	3.48 (.59)	2.67 (.67)	.81
	Grammar	3.46 (.62)	2.78 (.65)	.68	3.28 (.50)	2.73 (.66)	.55
	Vocabulary	2.91 (.60)	2.5 (.67)	.41	2.95 (.72)	2.75 (.64)	.20
	Total	16.74 (2.81)	13.19 (2.28)	3.55	16.39 (2.79)	13.79 (2.25)	2.60
Essay 4 - First Draft	Main Idea	3.61 (.67)	3.02 (.51)	.59	3.35 (.40)	3.13 (.84)	.22
	Organization	3.5 (.69)	3.59 (.75)	-.09	3.25 (.50)	3.6 (.66)	-.35
	Support	3.39 (.45)	2.83 (.68)	.56	3.28 (.41)	2.9 (.72)	.38
	Grammar	3.39 (.69)	2.83 (.60)	.56	2.9 (.55)	2.86 (.73)	.04
	Vocabulary	3.2 (.67)	2.76 (.54)	.44	3.2 (.41)	2.83 (.65)	.37
	Total	17.09 (2.14)	15.03 (1.65)	2.06	15.98 (1.23)	15.32 (2.28)	.66

Data in Table 1 also reveals that differences between rater scores and student's self-assessment scores observed in the first essays were reduced in both groups in the fourth essay except for vocabulary. Both groups saw the gap widen control (.03 for the control group and .17 for the experimental group). Organization was impacted the most moving from .72 on the first essay to -.09 on the second essay in the control group. Organization in the experimental group differed by .43 in the first essay, but in the second essay, students were giving higher scores with a differential of -.35.

Table 2 illustrates the independent t-tests that were used to examine the second research question, which investigates if significant differences existed between the rater scores and scores from the self-assessments for both groups on the first and fourth essays. In the first essay for the control group, all areas of writing were significant at the $p \leq .01$ level, with the exception of vocabulary (though still significant at the .05 level). In contrast, the experimental group had significant differences at the $p \leq .01$ level in two of the five areas of writing (support and grammar) as well as the combined score, and one area of writing at the $p \leq .01$ level (main idea).

The data reveals that the experimental group had a slightly higher level of agreement in their initial self-assessment of their essays. For the fourth essay, the gaps in the level of agreement for all areas of writing except for organization were still statistically significant. The experimental group saw a greater reduction in the number of significant discrepancies between their assessments and the rater assessments. Significant differences were reduced in main idea, support, grammar, and the combined score. By observing the data, it can be said that the experimental group made greater progress in scoring themselves in a manner comparable to the raters. Cohen's *d* was applied to calculate the strength of the effect. In reference to the effect sizess that Plonsky and Oswald (2014) recommend in second language research, the effect size was considered small (*d* = .39).

Table 2

Aggregate Level Analysis: T-test Results on First Drafts of Both Essays

		Control Group (*n*=23)			Experimental Group (*n*=20)		
	Area of Writing	*t*-ratio	df	p–value (two-tailed)	*t*-ratio	df	p–value (two-tailed)
Essay 1 - First Draft	Main Idea	-3.81	44	.0004**	-2.06	38	.0461*
	Organization	-3.53	44	.0010**	-1.80	38	.0795
	Support	-4.43	44	.0001**	-3.98	38	.0003**
	Grammar	-3.59	44	.0008**	-2.98	38	.0051**
	Vocabulary	-2.20	44	.0331*	-0.93	38	.3599
	Total	-4.69	44	.0000**	-3.21	38	.0027**
Essay 4 - First Draft	Main Idea	-3.33	44	.0018**	-1.08	38	.2869
	Organization	0.41	44	.6842	1.89	38	.0666
	Support	-3.31	44	.0019**	-2.02	38	.0500*
	Grammar	-2.97	44	.0048**	-0.24	38	.8079
	Vocabulary	-2.42	44	.0196*	-1.89	38	.0658
	Total	-3.67	44	.0006**	-1.07	38	.2899

Note: * significant at $p \leq 0.05$, ** significant at $p \leq 0.01$

Discussion

The overall purpose of this study was to evaluate the impact of peer-feedback on self-assessment, namely, whether peer interactions and commenting help students to minimize discrepancies that exist between rater assessments and their own. The first research question examined whether there were significant gaps between student self-assessments and the assessment by independent raters at the onset of the study. Generally, the results indicate that rater scores were far below the scores given by both groups of participants. In fact, significant gaps were observed in all the data except for organization and vocabulary in the experimental group. There are several possible explanations for the lack of compatibility found in the first essay. The most obvious reason is that the students in the study lacked experience in assessing their performance. In Japan, self-assessment is an uncommon practice in English instructional

contexts because the role of learner autonomy has not taken root. Iimuro and Berger (2010) note that the concept of learner autonomy has not been fully realized in Japan, nor has it been highly valued in the past. Another explanation for the inaccurate self-estimates, particularly the overestimation in writing performance, might be that students lacked an accurate measurement standard by which they assess themselves. Moreover, the level of proficiency may play a role in inaccuracy seen in the first essay, as previous studies found that low and intermediate level students tend to overestimate their language proficiency than more advanced proficiency students (Davidson & Henning, 1985; Blanche, 1988; Heilenman, 1990).

The second research question investigated if integrating peer feedback into the self-assessment process would help students to create self-assessment ratings that are more compatible with the assessment scores assigned by teachers or raters. A brief examination of the mean scores for both groups showed notable progress over the term. The only area of writing that did not see a reduction in the mean was vocabulary. This finding was observed in both groups, and can be attributed to the fact that vocabulary learning was not emphasized in the course. Other than identifying and defining unknown words encountered in the readings, vocabulary learning was incidental and unstructured. The reduction in many of the gaps observed at the outset of the study (in Essay 1) indicates that students seem to have improved their self-assessment skills through training and repeated exposure. It seems plausible that practice in self-assessment, coupled with teacher feedback, strengthened the agreement between the student and the rater assessments. This finding is in line with previous research by Mok, Ching, Cheng, Cheung, & Ng (2006) and Ross, Rolheiser, and Hogaboam-Gray (1999) that argue that continual training is a salient factor in developing the capacity to self-assess accurately.

The data shows an overwhelming improvement in the experimental group in reducing significant discrepancies found in the first essay. More specifically, gaps between student ratings and those given by the raters in main idea, support, grammar, and overall score were mitigated over the course. In contrast, other than organization, gaps were not significantly reduced in the control group. Several reasons seem plausible for this result. First, compatibility was achieved in the experimental group due to discourse generated in the peer feedback sessions, particularly when students were examining the models and rubrics. This discourse generated in the training sessions and particularly during peer review, seemed to develop a type of meta-language which was used when commenting on their peers' essays. Students started using the language on the rubrics and focused their attention on getting their peers to highlight significant parts of their essays, repeat main ideas to remind readers of what has been discussed, and clarify how new information relates to old information. This metalanguage facilitated the revision process and likely led to a greater awareness of one's writing ability. Second, students in the experimental group were likely to gain a more accurate perspective of their writing performance by reading numerous essays during peer work over the course. Greater accuracy can be attained when students evaluate their work after it is juxtaposed within a larger community of writers.

Conclusion

The findings in this study suggest the critical role that peer feedback can have on the process of self-assessment. The results indicate that lower-intermediate students can make more valid assessments of their work if peer support is provided. The results found that most students overestimated their written work in virtually every area of writing, but the level of agreement

between the raters and the students did improve over the course. Peer feedback made a more significant contribution to reducing discrepancies between the student self-ratings and the rater scores. Equally important is to train students to conduct peer review and self-assessment effectively by providing structure and transparency to the process. Another implication of this study is that despite the widespread acceptance of self-assessment as a valid tool in promoting autonomy, the small gains achieved after prolonged exposure indicate that self-assessment alone may not be enough for students to make accurate assessments of their work.

In view of the small sample size and narrow range of the participants, the results presented in this study need to be interpreted with caution. As far as self-assessment is concerned, one should bear in mind that the present study investigated first-year undergraduate students in an English writing course who have very little experience in writing academic English and being autonomous learners. Further research needs to examine the role peer feedback has on self-assessment by using a wider range of participants with varying levels of ability.

References

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System, 32*, 407-425.

Black, P. & Wiliam, D. (1998). Inside the Black Box: Raising Standards through Classroom Assessment. *Phi Delta Kappan*, (1), 1-13.

Black, P.; Harrison, C.; Lee, C.; Marshall, B; & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Berkshire, England: Open University Press.

Blanche, P. (1988). Self-assessment of foreign language skills: implications for teachers and researchers. *RELC Journal, 19*(1), 75-93.

Boud, D. & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education, 18*, 529-549.

Brew, A. (1999). Research and teaching: changing relationships in a changing context. *Studies in Higher Education, 24*(3), 291-301.

Cheng, W. & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing, 22*(1), 93-121. DOI: 10.1191/0265532205lt298oa

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing, 2*(2), 164-179.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, *24*, 331-350.

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education and Teaching International 32*, 175–187.

Heilenman, L.K. (1990). Self-assessment and placement: A review of the issues. In *AAUSC Issues in Language Program Direction: A Series of Annual Volumes*. Ed. Teschner, R.V.

Honsa, S. Jr. (2013). Self-assessment in EFL writing: a study of intermediate EFL students at a Thai University. *Voices in Asia Journal, 1*(1), pp. 34-57.

Hughes, I., & Large, B. (1993). Staff and peer-group assessment of oral communication skills. *Studies in Higher Education, 18*(3), 379–385.

Iimuro A, & Berger M. (2010). Introducing learner autonomy in a university English course. *Polyglossia 19*:127-141.

Jafarpur, A. (1991). Can naïve EFL learners estimate their own proficiency? *Evaluation and Research in Education, 5*, 145-157.

Kavaliauskiene, G. (2004). Quality assessment in teaching English for specific purposes. *ESP World*. Available: http://esp-world.info/Articles

Khodadady, E., & Khodabakhshzade, H. (2012). The effect of portfolio and self-Assessment on writing ability and autonomy. *Journal of Language Teaching and Research, 3*(3), 518-524.

Magin, D. & Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education, 26*(3), 288-297.

Min, H. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15*(2), 118-141.

Mok, M., Ching L., Cheng, D., Cheung, R., & Ng, M. (2006). Self-assessment in higher education: experience in using a metacognative approach in five case studies. *Assessment & Evaluation in Higher Education, 31*(4), 415-433.

Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay writing: A case study from geography. *Innovation in Education and Training International*, *32*(4), 324-335.

Nunan, D. (1996). Towards autonomous learning: Some theoretical, empirical and practical issues. In R. Pemberton, et al (Eds.), *Taking control: Autonomy in language learning*. pp. 13-26. Hong Kong: Hong Kong University Press.

Oldfield, K.A., & Mcalpine, J.M.K. (1995). Peer and self-assessment at the tertiary level: An experiential report. *Assessment and Evaluation in Higher Education*, 20, 125-132.

O'Malley, J. M., & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical approaches for teachers.* Boston, MA: Addison-Wesley Publishing Company.

Orsmond, P, Merry, S., & Reiling, K. (1997). A study in self-assessment: tutor and students' perception of performance criteria. *Assessment and Evaluation in Higher Education, 22*(4), 357-369.

Orsmond P., Merry S. and Reiling K. (2000). The use of student derived marking criteria in peer- and self-assessment. *Assessment & Evaluation in Higher Education, 25*(1): 23–38.

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*(4), 195-203.

Patri, M. (2002). The influence of peer feedback on self and peer-assessment of oral skills. Language Testing, 19(2), 109-131.

Peters, O., Körndle, H., & Narciss, S. (2018). Effects of a formative assessment script on how vocational students generate formative feedback to a peer's or their own performance. *European Journal of Psychology of Education, 33*, 117-143.

Plonsky, L. & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning 64*:4, 878–912.

Ross, J. (2006). The Reliability, Validity, and Utility of Self-Assessment. *Practical Assessment Research & Evaluation, 11*(10). Available online: http://pareonline.net/getvn.asp?v=11&n=10

Ross, J., Rolheiser, C. & Hogaboam-Gray, A. (1999). Effects of Self-Evaluation Training on Narrative Writing. *Assessing Writing, 6*(1), 107-132.

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing

classrooms. *Language Teaching Research, 8*(1), 31-54.
Sambell, K., McDowell, L., & Sambell, A. (2006). Supporting diverse students: Developing learner autonomy via assessment. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 158-168). London: Routledge.
Stefani, L. A. J. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69-75.
Sullivan, K., & Hall, C. (1997). Introducing students to self-assessment. *Assessment and Evaluation in Higher Education, 22*, 289-305.
To, J. & Panadero, E. (2019). Peer assessment effects on the self-assessment process of first-year undergraduates, *Assessment & Evaluation in Higher Education, 44*:6, 920-932.
Topping, K. (2003). Self- and peer-assessment in school and university: Reliability, validity and utility in M. Segers, F. Dochy and E. Cascallar (Eds). *Optimizing new modes of assessment: In search of qualities and standards* (pp. 55–87). Dordrecht, The Netherlands: Kluwer Academic Publishers.
Williams, E. (1992). Student attitudes towards approaches to learning and assessment. *Assessment and Evaluation in Higher Education, 17*, 45-58.

Appendix

The Training Session Procedure

	Control Group (*n*=23)	Experimental Group (*n*=20)
Training Session 1 (3 hours)	Negotiating the areas to be assessed, completing exercises on worksheets focusing on the areas of writing being rated (i.e. making clear topic sentences, organizing ideas, supporting main ideas with examples, creating different sentence structures with a clearly presented subject and verb, and using transitions and vocabulary presented in class), examining model essays, and assessing the five areas of writing using sample essays.	
Training Session 2 (2 hours)	Several sample essays were evaluated by students and the raters. Discrepancies in scores were compared and explained.	
	To familiarize students with the assessment criteria and the revision process, commenting and revision exercises were practiced individually on a short essay students wrote in a previous class.	To familiarize students with the feedback process, students examined a short essay from a peer partner using a checklist highlighting the issues to address during peer feedback and how to make helpful comments.
Editing 1st drafts	Checklists	Peer feedback
	Self-assessment in each area of writing using checklists and referencing model essays	
Teacher feedback	Essays were scored in the five areas of writing and written comments concerning improvement were made.	
Editing 2nd drafts	Same as above*	
Teacher feedback	Same as above*	

Note. *Not analyzed in the study