Perspectives

Classroom Self-Assessment—A Pilot Study

Dale T. Griffee
Seigakuin University

Student self-assessment is of great interest to teachers who want their students to take more responsibility for learning by judging their own progress. This exploratory study compares self assessment, teacher assessment and peer assessment in a Japanese university EFL class. Nineteen students gave oral presentations and each student rated her own performance in terms of eight categories (loudness, eye contact, etc.). The other students also assessed the talk, as did the teacher. The three types of assessment scores were added, averaged and then compared. The results suggest that student and teacher assessment scores were similar and the scores of the higher proficiency students were more similar to the teacher scores than the lower proficiency students' scores. There was no difference in the way the male and female students judged themselves, and the self-assessment scores tended to be similar to the teacher scores.

学習者に自らの英語能力や学習状況を評価させ、学習により積極的な責任を負わせたいと願う教師にとっては、学習者の自己評価の妥当性・信頼性は重要な問題である。本研究では、学習者の自己評価を、他の学習者が行なった評価、教師が行なった評価と比較した。まず。被験者である19名の学習者に英語で口頭発表させ、自らの発表を、声の大きさ、アイ・コンタクト等の8 項目で評価させた。同時に、同じ8 項目について、他の学習者にも評価させ、それらの評価得点の平均値を求めた。さらに教師も同様に評価を行なった。

これらの3 種類の評価を比較分析した結果、1) 学習者による評価が教師の行なった評価に近似していること、2) 英語能力の高い学習者は、英語能力の低い学習者に比べて、より教師に近い評価を行なったこと、3) 学習者による自己評価に男女差は認められなかったこと、が判明した。

学習者が自ら行なった評価は、教師の行なった評価と大きく異なることはなかった。

In many educational settings, a close relationship between assessment and curriculum has developed over the past twenty years (Fradd & McGee, 1994, p. 281), and it is now commonly accepted that the learner should have a role in classroom assessment (Griffee, 1995; LeBlanc & Painchaud, 1985; Nunan, 1988). Nevertheless, student self-assessment (SSA) is still not common in the field of teaching English as a second or foreign language. This report presents the results of a

limited investigation of the effectiveness of self-assessment of an oral presentation activity in a Japanese university EFL classroom compared with peer-assessment and teacher assessment.

Classroom Research on the Use of Learner Self-Assessment

Self-assessment is also known as self-report, self-rating or self-evaluation and has been defined as checking one's own performance on a learning task after it has been completed (Richards, Platt, & Platt, 1992, p. 327). Wesche, Paribakht and Ready (1996, p. 199) state that "self-report procedures usually require candidates to rate their ability to do certain things using their L2, or their knowledge of particular elements or patterns of the L2."

Current trends now favor communicative language teaching. This pedagogy brings the learner to center stage (Graves, 1996, p. 24) and supports autonomous learning and the learner-centered classroom, formats which favor the use of SSA. For example, Dickinson (1993, p. 330) lists five characteristics of an autonomous learner: The autonomous learner can identify what has been taught, can formulate his own learning objectives, can select and implement his learning strategies, and can self-assess. In discussing the learner-centered classroom, Nunan (1988, p. 116) argues that both the learner and the teacher should be involved in evaluation, and Griffee's review (1995, p. 3) identifies SSA as an important characteristic of learner-centered classrooms.

Proponents of SSA offer wide-ranging justifications for its use, some of which are supported by empirical studies and some of which remain working hypotheses. These can be reduced to nine general arguments.

1. Self-assessment raises self-consciousness by focusing learner attention on performance (Nunan, 1988, p. 116; Oskarson, 1989, p. 4).
2. Self-assessment increases learner motivation (Rolfe, 1990, p. 169); a review of the literature (Blanch, 1988, p. 82) cites eight studies supporting this suggestion.
3. Self-assessment promotes learning by giving learners training in evaluation (Oskarson, 1989, p. 3). This occurs when learners address questions such as "What am I learning?" and "How well am I learning?"
4. The criteria for self-assessment can be directly related to course goals and objectives allowing the learner to better understand course organization (Brindley, 1989, p. 60).
5. Self-assessment can result in learners becoming more goal-oriented (Rolfe, 1990, p. 169), thereby exerting more effort to achieve their

goals, and even formulate goals themselves (Oskarson, 1989, p. 4). Within the context of given course objectives, SSA can show both learner and teacher new ways to accomplish those objectives (Legutke & Thomas, 1991, p. 243).

6. Self-assessment can help learners identify preferred materials as well as learning styles and strategies (Nunan, 1988, p. 130).
7. Self-assessment helps promote a cooperative classroom (Brindley, 1989, p. 60).
8. Self-assessment frees the teacher from being the only person in the classroom concerned with evaluation (Brindley, 1989, p. 60; Oskarson, 1989, p. 4; Rolfe, 1990, p. 169).
9. Self-assessment can continue after the course is finished. This is an important consideration since no single teacher or course can teach the entirety of a language. Therefore, learners must continue to acquire language through their own effort (Dickinson, 1987, p. 136; Oskarson, 1989, p. 5).

On the other hand, there have also been objections to wide-spread use of SSA. These can be summarized by the following three arguments. The first is that many learners lack the ability to self-assess and cannot do it reliably (Oskarson, 1989, p. 2). Citing Blanch and Merino (1989), Cohen (1994, p. 199) lists five factors that can threaten the validity of self-assessment, including the fact that learners may not be able to accurately report or assess what is often subconscious behavior. Second, learners may lack motivation to self-assess because of culturally-based expectations of appropriate classroom behavior and activities (Cohen, 1994, p. 199; Lynn, 1995, p. 37). Additional problems come from subjectivity and the natural desire of students to inflate their ratings, whether this is intentional or not (Brindley, 1989, p. 61; Dickinson, 1987, p. 134). A third obstacle to SSA is the lack of shared valid criteria for the learners and the teacher to use in assessment (Blanch, 1988, p. 82; Cohen, 1994, p. 199). This situation occurs when the teacher asks student to assess their work without clearly explaining the criteria which must be used. The lack of learner training in assessment (Cohen, 1994, p. 199) is related to this lack of criteria and probably results from unwarranted teacher assumptions that learners have the tools for self-assessment (LeBlanc & Painchaud, 1985, p. 675).

Such objections account for teacher skepticism (Brindley, 1989, p. 60) and, when combined with the natural fear of change (Rojas, 1995, p. 32), may account in part for the lack of SSA in many classrooms today.

However, many of these objections are based on teacher supposition rather than actual research findings. For example, a study using confirmatory factor analysis and a multitrait-multimethod design (Bachman & Palmer, 1989, p. 22) reports that self-ratings can be a reliable and valid measure of communicative language ability.

Regarding the question of consistent agreement between individual self-assessments and other sources, a review of 16 articles (Blanch, 1988, p. 81) reported a pattern of consistent agreement between SSA and a variety of external criteria. However, other research findings are less positive. A study of adult learners of various linguistic backgrounds in Australia (Rolfe, 1990, p. 177) reported that students consistently rated themselves lower than their peers' ratings. Whereas Dickinson (1987, p. 150) suggested that learners are biased in their own favor, Rolfe (1990, p. 178) concluded that learners are more critical of themselves than their teachers are; thus SSA was not a reliable indicator of oral ability as compared to teacher-assessment (TA). In comparing SSA to peer-assessment (PA), Rolfe reported that the PA may therefore be more reliable. Falchikov and Boud (1989, p. 398) investigated whether fourth year university students were more accurate in their SA than first year students and concluded that they were not. This is in accord with the findings of Griffee (1996, p. 32), who reported on a classroom SSA project in which there was no major difference in self-evaluations among first-year, second-year, and third-year oral conversation classes at a Japanese university. Relative to possible differences in male and female responses to self-assessment, Falchikov and Boud (1989, p. 396) concluded that gender differences are under-researched and that no conclusions can be drawn. They also question whether learners overestimate or underestimate themselves relative to teacher assessment, and stress the need for further research investigating the reliability of self-assessment among different groups of learners as well as the development of methods to improve the learners' ability to accurately estimate their performance.

The Study

Research Questions

The purpose of this exploratory study is to examine the operation of SSA in a Japanese university EFL classroom setting. The specific research questions are:

1. To what extent will SSA, PA, and TA test scores agree?
2. Will there be a higher level of agreement between more proficient

students and the teacher than between less proficient students and the teacher?

3. Will there be any gender differences in self-assessment?
4. Will SSA be higher or lower than TA?

Methods

Subjects: The students who participated in this study were enrolled in the second semester of a first-year required English oral conversation course at a small liberal arts university in Japan. The total class enrollment was 24, with 12 females and 12 males, but only 19 students were present during the two class periods when the study was conducted. The majority of the students were 18 or 19 years old. The subjects' Secondary Level English Proficiency (SLEP®) test scores averaged 42.0, which is equivalent to 400 on the TOEFL®. The SLEP® test scores were used to divide the students into high-proficiency and low-proficiency groups in the following way: The four subjects with scores of the mean value 42 were eliminated, leaving 10 students with scores over 42, eight of whom gave oral presentations and 10 students with scores under 42, seven of whom gave oral presentations. The presentation theme for all students was "How I study vocabulary."

Materials: A short score sheet (see the Appendix) was constructed which asked students to evaluate each oral presenter on eight points within three categories—voice, body language, and content. Under the category of "voice," the points to be rated were loudness, clarity, and speed; under "body language," the points were eye contact and gestures; under "content," the points were introduction, interesting talk, and conclusion. Each point could be rated on a Likert-type scale with values from one to three, with three as the highest score.

Procedures: A 45-minute training session was conducted by a Japanese native speaker and an English native speaker. Each category was explained in some detail in both Japanese and English, then each of the eight evaluation points was illustrated by the English native speaker in all three conditions and discussed by the Japanese native speaker.

The students were then assigned the oral presentation topic and two class sessions were spent making the oral presentations. When making the oral presentation, the student came to the front of the room and stood behind the teacher's desk. The talk had no time limit, although most talks were completed in under five minutes. After the oral presentation, the teacher, the student giving the talk, and the rest of the students completed their score sheets.

Analysis

Pearson product-moment correlations were used to analyze the individual self-assessment, the PA, and the TA scores, with the alpha level set at .05. Use of the Pearson correlation procedure assumes the presence of interval scales, equivalent reliability, independent data, a normal distribution, and a linear relationship (Hatch & Lazaraton, 1991, p. 549). To check these assumptions, descriptive statistics were generated by StatView 4.5 for the Macintosh (1992). Correction for attenuation[1] was done using the formula from Guilford and Fruchter (1973, p. 439). The non-parametric Wilcoxon Signed Rank Test was also used to determine if there was any difference between the SSA scores and teacher scores. Cronbach's alpha, a measure of reliability, and the standard error of measurement (SEM) were calculated on a spreadsheet from the formula provided in Brown (1996, p. 196).

Results

The descriptive statistics reveal similarities between the SA and the TA scores (Table 1), with a mean assessment score of about 1.8 for each group. However, the mean PA score of 2.28 was higher than both SA and TA scores. The SLEP® scores formed a fairly normal distribution. Therefore, a Pearson correlation was calculated for both groups of students between their SA scores and the teacher scores to determine which group's ratings was closest to the ratings of the teacher. The correlation between the higher proficiency students' scores and the teacher scores

Table 1: Descriptive Statistics, Alpha Reliability, and SEM for SSA, PA, and TA

	SSA	PA	TA
Mean	1.85	2.28	1.80
Standard Deviation	.63	.34	.74
Minimum	1.00	1.20	1.00
Maximum	3.00	3.00	3.00
Median	2.00	2.30	2.00
Skewness	.12	-.62	.33
Kurtosis	-.53	.63	-1.10
Chronbach's alpha	.84	.77	.79
SEM	1.12	.56	1.63

was .241 ($p < .0547$), whereas the correlation between the lower proficiency students' scores and the teacher scores was .187 ($p < .695$). To determine whether there was a significant difference between all SSA scores and TA scores, a Wilcoxon Signed Rank Test was performed. The results ($z = -.575$, $p < .5653$) indicate that there was no significant difference between the two sets of scores.

Pearson correlations between the total scores for student assessment, PA, and TA were calculated and corrected for attenuation (Table 2). A low correlation was found between SSA and TA, a slightly higher correlation was found between SSA and PA, and a relatively strong correlation was found between PA and TA. R square, which is the Pearson correlation coefficient squared and expressed as a percentage, gives an indication of the magnitude of the relationship. The figure of six percent for the relationship between the SSA scores and the teacher assessment scores indicates that only six percent can be accounted for by the correlation, whereas 13% of the relationship between SSA and PA is explained, and 42% of the relationship between SSA and TA is accounted for by the correlation, as shown below.

To investigate the existence of gender differences in assessment score values, the scores were totaled for each student and the number of student scores that were higher and lower than TA scores was counted (Table 3). To account for standard error, if the difference between higher than TA and lower than TA scores was plus or minus one, these values were eliminated and the resulting scores are referred to as adjusted scores.

There were 12 students who rated themselves higher than the teacher's ratings, and seven students who rated themselves lower. After eliminating the scores with values of plus or minus one from the teacher's scores, there were ten students who rated themselves higher than the teacher and six students who rated themselves lower. Of the ten who rated themselves

Table 2: Pearson Product-Moment Correlations (r) Between SSA, PA and TA

	r	p	C/A	R^2
SSA and TA	.207	.0104	.254	.06
SSA and PA	.285	.0003	.354	.13
PA and TA	.508	.0001	.651	.42

SSA = student self-assessment, TA = teacher assessment, PA = peer assessment, C/A = correction for attenuation

Table 3: Individual Student Scores Higher than TA and Lower than TA

	Higher	Lower	Adjusted Higher	Adjusted Lower
Males	7	3	5	3
Females	5	4	5	3
Totals	12	7	10	6

higher, five were males and five were females. Of the six who rated themselves lower, three were males and three were females. Thus, there were no gender differences in scoring in the restricted sample used here.

Discussion

The first research question asked whether the SSA, PA, and TA test scores agreed. The descriptive statistics show that the SSA scores were similar to the TA scores. The correlations in Table 1 indicate a low correlation between the SSA and TA, a modest agreement between SSA and their peers, and a higher agreement between PA and TA. On the face of it, this would seem to suggest that students did not agree with the teacher in their assessment of themselves, whereas, as a group evaluating each other (PA), their scores were similar to their teacher's scores. However this result should be interpreted cautiously. The SSA and teacher scores suffered from restriction of range, suggesting that the correlation coefficients were very likely depressed. The use of a limited Likert scale, with values of only one to three, produced the low variance. The relationship between SSA and TA therefore requires further investigation using a larger number of subjects and an instrument with a greater number of choices, permitting more variance.

The second research question asked whether higher proficiency learners would exhibit better agreement between their self-evaluations and the teacher evaluations than the lower proficiency group. The answer to this question was inconclusive. The correlation between the teacher scores and the higher ability students scores ($r = .241$; $p < .05$) was higher than the correlation between the lower ability students and the teacher ($r = .187$; $p < .70$), but was not statistically significant.

The third research question involved the impact of gender on the evaluation process. As shown in Table 3, the number of male students who rated themselves higher or lower than the teacher was exactly the same as the number of female students who scored themselves higher

or lower. In this limited study, gender was not significant, but it should be noted that the number of subjects was low.

Research question four asked whether the SSA scores would be higher or lower than the teacher scores. The results indicate there was no difference between SSA scores and teacher scores. This suggests that students were assessing themselves in a manner similar to the teacher and provides some support for the validity of SSA, keeping in mind the limitations of this pilot study.

Conclusion

Problems with the present study include the restricted Likert scale which produced a narrow band of scores, the small number of subjects, and the use of a data collection instrument which was not validated. Therefore the findings reported here are not generalizable. Nevertheless, this preliminary study is encouraging in that the student peer-assessment appears to be similar to teacher assessment in the group studied. Suggestions for future research include use of a validated data collection instrument, a much larger number of subjects and a five-point Likert scale to increase the score range. There is also a clear need for longitudinal studies which examines the effect of experience and training on student assessment.

Acknowledgments

Thanks are due to Sonia Yoshitake for help in the training session and J. D. Brown for help in calculating the correction for attenuation.

Dale T. Griffee teaches at Seigakuin University. His major research interests are curriculum and evaluation and he has edited (with David Nunan) *Classroom Teachers and Classroom Research* in the JALT Applied Materials series. He can be contacted at Seigakuin University, 1-1 Tosaki, Ageo-shi, Saitama-ken 362 or by e-mail: <Dale_Griffee@ringo.net>.

Notes

1. Attenuation is a correction for reliability applied to a correlation coefficient. Correlation assumed perfect reliability. If the reliability is .70, this means that 30% of the score is error which lowers the correlation coefficient. Attenuation takes this into account.

References

Bachman, L., & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*(1), 14-29.

Blanch, P. (1988). Self-assessment of foreign language skills: Implications for

teachers and researchers. *RELC, 19*(1), 75-87.
Blanch, P., & Merino, B. (1989). Self-assessment of foreign language skills: Implications for teachers and researchers. *Language Learning, 39*(3), 313-340.
Brindley, G. (1989). *Assessing achievement in the learner-centred curriculum.* Sydney: National Centre for English Language Teaching and Research: Macquarie University.
Brown, J.D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.
Cohen, A. (1994). *Assessing language ability in the classroom.* Boston, MA: Heinle & Heinle.
Dickinson, L. (1987). *Self-instruction in language learning.* Cambridge: Cambridge University Press.
Dickinson, L. (1993). Aspects of autonomous learning: Interview with T. Hedge. *English Language Teaching Journal, 47*(4), 330-336.
Falchikov, N. & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395-430.
Fradd, S., & McGee, P. (1994). *Instructional assessment: An integrative approach to evaluating student performance.* Reading, MA: Addison-Wesley.
Guilford, J.P. & Fruchter, B. (1973). *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill.
Graves, K. (1996). A framework of development processes. In K. Graves (Ed.), *Teachers as course developers.* (pp. 12-38). Cambridge: Cambridge University Press.
Griffee, D.T. (1995). Implementation of student originated goals and objectives in a learner-centered classroom. *The Language Teacher, 19*(12), 14-17.
Griffee, D.T. (1996). A longitudinal study of student feedback: Self-assessment, course evaluation, and teacher evaluation. *Temple University Japan Research Studies in TESOL, 3,* 27-39.
Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* Boston, MA: Newbury House Publishers.
LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly, 19*(4), 673-687.
Legutke, M., & Thomas, H. (1991). *Process and experience in the language classroom.* London: Longman.
Lynn, M.J. (1995). Caveat emptor: Using innovative classroom assessment. *TESOL Journal, 5,* (1), 36-37.
Nunan, D. (1988). *The learner-centred curriculum.* Cambridge: Cambridge University Press.
Oskarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing, 6*(1), 1-13.
Richards, J., Platt, J., & Platt, H. (1992). *Longman dictionary of language teaching & applied linguistics.* London: Longman.
Rojas, V. (1995). A higher education: Practicing what you preach in teacher education. *TESOL Journal, 5*(1), 32-35.
Rolfe, T. (1990). Self-and-peer-assessment in the ESL curriculum. In G. Brindley

(Ed.), *The second language curriculum in action* (pp. 163-186). Sydney: National Centre for English Language Teaching and Research: Macquarie University.

StatView 4.5 [Computer software]. (1992). Berkeley, CA: Abacus Concepts.

Wesche, M., Paribakht, T., & Ready, D. (1996). A comparative study of four ESL placement instruments. *Performance testing, cognition and assessment: Selected papers from the 15th language testing research colloquium, Cambridge and Arnhetm*. Cambridge: Cambridge University Press.

(Received January 24, 1997; revised May 2, 1997)

Appendix

Oral Presentation Score Sheet Used by Students and Teacher

Speaker _____ Date _____

	needs work	ok	great
VOICE			
loudness	1	2	3
clear	1	2	3
speed	1	2	3
BODY LANGUAGE			
eye contact	1	2	3
gesture	1	2	3
CONTENT			
introduction	1	2	3
interesting talk	1	2	3
conclusion	1	2	3