Rater Training Effects on Peer Assessment of EFL Individual Presentations : An Interim Report

Hidetoshi SAITO

Contents

An increasing desire to integrate instruction and assessment has directed language teachers towards making use of alternative assessments. One of the alternative assessments which draws attention is peer assessment (Brown, 1998; O'Mally & Valdez Pierce, 1996). With peer assessment, students assess their peers' performance in language classrooms. This classroom assessment carries some advantages over other assessments: (1) receiving feedback from multiple sources enhances self-awareness; noticing the gap between self- and others' perception motivates for change. (London & Tornow, 1998); (2) responsibility for managing the assessment leads to increasing responsibility for learning; (3) evaluating peers sensitizes students to the evaluation criteria (Brown, 1998); and (4) discussing individual strengths and weaknesses encourages connectivity in the learning community.

There has been a concern about the biases of peer assessment, however. Anecdotal evidence suggests that language teachers have been hesitant to bring peer assessment into classrooms because of the lack of reliability in student rating. This is very unfortunate given the fact that teachers would miss a chance of making most of the advantages of peer assessment described above. One strategy to get around this problem is to give rater training to students. The present study examines the effects of rater training in assessing EFL individual presentations.

Literature Review

Both quantitative and qualitative reviews of peer assessment have suggested consistent evidence of predictive and concurrent validity of peer assessment in school and workplace settings (Falchikov & Goldfinch, 2000; Fletcher & Baldry, 1999; Harris & Schaubroeck, 1998; Kane & Lawler, 1978; O'Donnelle & Topping, 1997; Topping, 1998). However, studies have reported several biases of peer assessment including friendship bias (Falchikov, 1995) reference bias (Holzbach, 1978) purpose bias (Battenhausen & Fedor, 1997); collusive bias (O'Donnelle & Topping, 1997); feedback bias (DeNishi et al., 1983). But these biases are typical of most human rating (Fletcher & Baldry, 1999), and efforts need to be made in order to minimizing biases by (1) setting clear rating criteria; (2) informing raters of the goals and limits of peer assessment; and (3) familiarizing raters with the instrument (Edwards & Ewen, 1996).

The best way to achieve this goal probably is to give students rater training sessions. In general, rater training should reduce extreme rating and enhance within-rater consistency. This does not mean training could avoid variability across individuals, but help raters to become self-consistent. In a study of training professional raters of an occupational English test in Australia, Lumley and McNamara (1995) suggested that differences in rater severity remain after training. Literature in psychology and education has demonstrated positive effects of rater training and many argue for the significant role it plays in performance assessment and appraisal (e.g., Martin & Bartol, 1986).

In language testing literature, studies on rater training are very limited and it was not possible to locate even a single study that directly addressed the question of rater training effects in student peer assessment. There are studies, however, which surely give some insight into rater training of student peer assessment in language learning classrooms. Patri (2002) examined the effects of peer feedback on individual presentations. The treatment group repeatedly received peer feedback and demonstrated more convergence with teacher rating than the control group. In a study by Stanley (1992), a seven hour training on peer-commenting on EFL essays resulted in more comments from the treatment group than from the control group. Although these two studies may bolster an assumption that feedback and training may lead to any improvement in student peer assessment, evidence is premature and some questions still remain unanswered. It is not clear, for example, whether or not rater training does have impact on peer assessment of individual presentations. If so, to what degree does peer assessment by those receiving rater training converge with teacher rating? Does the trained group become more severe in rating? Little is understood, in addition, as to the effects of a rater training session on students' attitudes toward peer assessment. An assumption is that those who receive rater training should feel more positive about peer assessment than those who don't. The following research questions were thus formulated:

(1a) Is the instructors' rating more similar to the treatment group's peer rating than to the control group's? As a corollary of this research question, the following is also asked:

(1b) Is the treatment group's rating more severe than the control because of the training?

(2) Is the treatment group's attitudes toward peer assessment more positive than the control's?

Method

Participants and Setting

Seventy-eight university freshmen (ages 18-19; 55 males and 23 females) of economics and information and management majors participated in the study. This school is a small liberal arts university located in Sapporo, Japan. The participants were all ethnic Japanese and their SST (Standard Speaking Test) levels ranged from levels two to three. They enrolled in three sections of an English communication course as part of their requirements. This course met twice a week for 13 weeks of a semester. Three Japanese teachers of English also participated in the study as the raters. All three sections were taught by one of the raters using identical instruction materials.

Procedure

As part of the course requirements, each student was asked to give an individual presentation. Students chose their topic of interest such as travel, hobbies, and club activities. All the students received instruction on the aspects of presentation that would be in peer assessment and had practices in class (Table 1). These practices were held over 5 class sessions.

Table 1 Schedule of practice and assessment

Sessions	Tasks (Focused Aspects of Presentation)
1	Information gap (Gestures/Posture)
2	Group mock presentation (Eye contact/Visual Aids)
3	Group mock presentation (Introduction/Body/Conclusion/Purpose)
4-5	Class mock presentation (Diction/Pace/Intonation/Grammar/Vocabulary)
5	Rater training session for treatment group and an irrelevant writing task for control group
6	Student-teacher conference on draft
7-9	Presentations/Peer assessment
10	Feedback/Attitude questionnaire

In the fifth session, each section was randomly divided into the treatment and control groups. While the control group was assigned an irrelevant writing task, the treatment groups moved to a different room and received a forty minute-long rater training session. Although forty minutes is not very long, such a short

training session should simulate the time most English classes in Japanese institutional settings could spare, since they often suffer from a chronic lack of instruction time. In the training session, the instructor first explained each item in detail. By this time, students were believed to be familiar with the features of presentation that are reflected in the items, because over the five sessions prior to training, all the students received instruction that focused on each presentation feature (Table 1). Students viewed three video-taped presentations of former students and rated each presentation. They were asked to compare in groups their own ratings with those of other students. Students then reported by raising their hands what rating was given to each item, and compared their ratings to the teacher's rating. The teacher explained why a certain rating was more appropriate for a certain presentation and pointed out and discouraged obvious over- and under-ratings.

Instruments

The instrument used in the study was adapted from Yamashiro (1999). This instrument contains 13 items accompanying a four-point rating scale (see Appendix). She reported an inter-rater reliability of .69 and interclass reliability of .97 for this instrument. One item (Visual Aids) was added for this study because of the instructional focus in the lessons.

An attitude questionnaire contained four items that asked student attitude toward peer assessment (see Table 9). Questions involved the fairness of using peer assessment in grading, trust in peer rating, and learning by assessing peers.

Analysis

The Rasch analysis was employed for checking dimensions and generating item difficulty measures, presentation quality measures, and rater severity measures. To examine the magnitude of similarities between peer rating and teacher rating, Pearson correlations of quality measures were calculated (Research question 1a). For Research question 1b, a MANOVA on rater severity measures by treatment and control was implemented in order to see any difference between the two groups. A MANOVA on attitude measures was also done for testing any difference between control and treatment (Research question 2).

Results

The Rasch Analysis of Items

Table 2 shows the results of the initial Rasch analysis of item difficulty measures. If one examines the fit statistics of the present results, two items seem to fall into the misfitting item category based on Wright and Linacre's (1994) criteria. That is, Visual Aids and Gestures satisfy both conditions of z values exceeding 2.0 and infit/outfit statistics deviating from the range of .4 to 1.2.[(1)] In addition, Eye Contact is fairly close to

meeting these criteria. Statistically speaking, these items seem to measure a different trait from the one the rest of the items measure. Conceptually, these three items share the same "visual" aspects of presentation in addition to Posture, which is not misfitting. A decision was made, then, that these four items be analyzed separately from the rest of the items which now share only verbal aspects of presentation.

Table 3 and 4 show the separate Rasch analyses of item difficulty measures for verbal skill items and visual skill items. None of these items is misfitting except for Diction in Table 3. It was decided, however, to keep this item in the subsequent analyses because (1) diction is an important aspect of presentation to be evaluated and (2) it is often the case that once some misfitting items are removed, other items become misfitting. Table 3 and 4 also suggest that both sets of items attain high reliabilities.

Table 2 The Initial Rasch Analysis of Item Difficulty Measures

		Infit		Outfit			
M	SE	MSQ	z	MSQ	z	pbs	Items
-.40	.04	0.8	-6	0.8	-6	.31	1 Pace
.10	.04	0.7	-9	0.7	-8	.36	2 Intonation
.01	.04	0.9	-1	1.0	0	.31	3 Diction
-.01	.04	0.9	-4	0.9	-3	.35	4 Posture
1.09	.03	1.2	5	1.2	5	.41	5 Eye Contact
1.91	.03	1.6	9	1.6	9	.35	6 Gesture
-.62	.04	0.8	-6	0.8	-5	.30	7 Introduction
-.58	.04	0.8	-6	0.8	-6	.31	8 Body
-.42	.04	0.7	-9	0.8	-7	.34	9 Conclusion
-.53	.04	0.8	-5	0.9	-4	.29	10 Language Use
-.52	.04	0.8	-4	0.9	-3	.28	11 Vocabulary
-.28	.04	0.7	-9	0.7	-9	.40	12 Purpose
.26	.04	1.8	9	1.8	9	.42	13 Visual Aids

Separation 19.15 ; Reliability 1.00
Notes. M = item difficulty measures ; SE = standard error ; MSQ = mean square ; z = z-value ; pbs = point biserial correlation.

Table 3 Item Difficulty Measures of Verbal Skills

		Infit		Outfit			
M	SE	MSQ	z	MSQ	z	pbs	Items
-.06	.05	0.9	-2	0.9	-2	.38	1 Pace
.65	.04	1.0	0	1.0	0	.36	2 Intonation
.53	.04	1.4	9	1.4	9	.31	3 Diction
-.35	.05	0.9	-3	0.9	-3	.39	7 Introduction
-.31	.05	1.0	-1	1.0	-1	.34	8 Body
-.09	.05	0.8	-4	0.8	-4	.40	9 Conclusion
-.24	.05	0.9	-1	0.9	-1	.36	10 Language Use
-.23	.05	1.0	0	1.0	0	.35	11 Vocabulary
.10	.05	1.1	1	1.1	1	.34	12 Purpose

Separation 7.52 ; Reliability .98
Notes. M = item difficulty measures ; SE = standard error; MSQ = mean square ; z = z-value ; pbs = point biserial correlation.

Table 4　Item Difficulty Measures of Visual Skills

		Infit		Outfit			
M	SE	MSQ	z	MSQ	z	pbs	Items
-.74	.03	.03	0	1.1	3	.28	4 Posture
.22	.03	.03	-4	0.9	-4	.47	5 Eye Contact
1.02	.03	.03	2	1.0	0	.49	6 Gesture
-.51	.03	.03	1	1.0	0	.51	13 Visual Aids

Separation 7.52 ; Reliability .98
Notes. M = item difficulty measures ; SE = standard error; MSQ = mean square ; z = z-value ; pbs = point biserial correlation.

Research Question（1a）Is the instructors' rating more similar to the treatment group's peer rating than to the control group's?

Table 5 displays the descriptive statistics of presentation quality measures of the three rater groups. One noteworthy point here is that while teachers demonstrate much wider ranges of rating on verbal skills, student ratings are as wide a range as teachers' on visual skills. Table 6 shows Pearson correlations of teacher and peer ratings on presentation quality measures. The results suggest that both treatment and control groups correlate highly with teachers in assessing both skills, and the correlation of assessments of visual skills are higher than verbal skills. Note also that the control group's correlations are slightly higher than those of treatment.

Table 5　Descriptive Statistics of Presentation Quality Measures

		N	Min	Max	Mean	SD
Verbal	Teachers	3	-1.44	3.59	.99	1.00
			2.60	3.80	3.20	.27
	Treatment	36	.27	3.28	1.12	.56
			2.70	3.60	3.08	.14
	Control	38	-.04	2.05	.86	.53
			2.70	3.40	3.02	.17
Visual	Teachers	3	-2.20	3.43	.00	1.10
			1.30	3.70	2.19	.57
	Treatment	36	-2.23	3.32	.24	.47
			1.60	3.80	2.72	1.00
	Control	38	-2.61	1.96	-.10	1.00
			1.40	3.40	2.56	.51

Note. The upper rows show logit scores based on a Rasch analysis; the lower rows show numbers based on raw scores.

Table 6　Correlations of Presentation Quality Measures

	Verbal		Visual	
	Treatment	Control	Treatment	Control
Teacher	.627*	.658*	.833*	.891*

* = significant at .01

Research Question (1b) Is the treatment group's rating more severe than the control's because of training?

Table 7 shows the descriptive statistics of rater severity measures, and the results of a one-way MANOVA appear in Table 8. As seen, there is no statistically significant difference between control and treatment groups, suggesting that there were no effects of rater training on rater severity measures.

Table 7 Rater Severity Measures of Control and Treatment Groups

		Mean	SD
Verbal skills	Control	-1.21	1.24
	Treatment	-.91	.95
Visual skills	Control	-.28	.79
	Treatment	-.11	.57

Table 8 MANOVA on Rater Severity Measures

	Value	F	Hdf	Error df	sig.	Eta^2
Wilk's λ	.982	.644	2	71	.528	.018

Research Question (2) Is the treatment group's attitudes toward peer assessment more positive than the control?

Table 9 shows the descriptive statistics of student attitudes toward peer assessment. The overall trend of student responses is in a positive direction, since all the means exceed 2.5 on 4 point scales. There is not much difference in means between the two groups, and such is confirmed by a one-way MANOVA, Wilk's $\lambda = .865$, $F = 1.756$ (Hdf = 4, Edf = 45), $p > .05$, $eta^2 = .135$.

Table 9 Descriptive Statistics of Student Attitudes toward Peer Assessment

	Control		Treatment	
Items	M	SD	M	SD
1) It is fair that peer assessment results are incorporated into my grade.	2.72	.81	3.00	.67
2) By evaluating peers' presentations, I learned.	3.30	.61	3.13	.63
3) I believe that my classmates evaluated others fairly.	2.85	.66	3.09	.73
4) As raters, peer students are reliable.	3.19	.74	3.17	.83

Discussion and Implications

The following summarizes the results of the present study:

(1) Based on a Rasch analysis, two aspects or dimensions of individual presentation skills are proposed, verbal and visual presentation skills.

(2) Presentation quality measures by both treatment and control groups correlated highly with those of teachers. For both groups, in particular, evaluation of visual skills correlated more highly than did evaluation of verbal skills.

(3) There was no statistically significant difference between treatment and control groups on rater severity measures.

(4) There was no difference between treatment and control groups on attitude measures. Both groups, however, showed overall positive attitudes toward peer assessment

These results put forward several implications for practitioners and researchers. First, the results of the present study suggest that a rater training session may not be useful if it is administered in the same way as the present study did, in a rather minimal forty minute session. The short length of the training session, however, may not be the only reason why the treatment group did not reap the benefits from the training. As shown in Table 1, both the control and treatment groups underwent a series of classroom activities that raised student awareness of the significance of the aspects of presentation which are all reflected in each item. Chances are these activities are sufficient for students to learn to assess these aspects reasonably well and attain a fairly high convergence with teacher rating. This possibility can be implied in, regardless of groups, rather high correlations with the teacher ratings in Table 6. This being the case, teachers could better use the time for classroom activities rather than for a forty minute rater training session.

Another possibility is that it may take a longer training time to cultivate peer-rating skills, in particular, rating of verbal skills. Note that the correlations of verbal skill ratings are slightly lower than those of visual skill ratings (Table 6). Future study should examine whether or not a longer training session actually has any impact on subsequent peer rating.

Although the results of the present study indicate that there was no effect of rater training, both control and treatment groups' ratings demonstrated fairly high convergence with teacher ratings as well as positive attitudes toward peer assessment. In fact, this is a very encouraging result for those teachers who doubt the utility of peer assessment in language classrooms and fear the unreliability of student rating.

Notes

(1) Outfit statistics (unweighted mean square) are the squared standardized residuals for the item averaged over the number of test-takers. Outfit statistics are sensitive to unexpected responses to the item at far above or below a persons' ability. Infit statistics (weighted mean square) are the squared residuals weighted by the variance. Infit statistics are sensitive to unexpected responses to the item around the person's ability. Note that z values derive from standardizing outfit/infit mean squares. See Wright and Masters (1982) for more details.

Acknowledgement

This research was in part supported by an Education and Science Ministry of Japan Research Grant awarded to the author (task no. 13780147).

REFERENCES

Bettenhausen, K. L., & Fedor, D. B. (1997). Peer and upward appraisals--a comparison of their benefits and problems. *Group and Organization Studies*, 22(2), 236-263.

Brown, J. D. (Ed.). (1998). *New ways of classroom assessment*. Alexandria, VA: Teachers of English to the Speakers of Other Languages.

DeNishi, A., Randoloph, W. A., & Blencoe, A. G. (1983). Potential problems with peer ratings. *Academy of Management Journal, 26,* 457-464.

Edwards, M. R., & Ewen, A. J. (1996). 360^{0} *feedback: The powerful new model for employee assessment & performance improvement.* New York, NY: AMACOM.

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

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322.

Fletcher, C., & Baldry, C. (1999). Multi-source feedback systems: A research perspective. *International Review of Industrial and Organizational Psychology, 14*, 149-193.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41,* 43-62.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology, 78,* 579-588.

Kane, J., & Lawler, E. E. I. (1978). Methods of peer assessment. *Psychological Bulletin, 85,* 555-586.

London, M., & Tornow, W. W. (1998). 360-degree feedback--More than a tool! In W. W. Tornow & M. London & C. Associates (Eds.), *Maximizing the value of 360-degree feedback: A process for successful individual and organizational development* (pp. 1-8). San Francisco, CA: Jossey-Bass.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

Martin, D. C. & Bartol, K. M. (1986). Training the raters: A key to effective performance appraisal. *Public Personnel Management, 15*(2), 101-109.

O'Donnell, A. M., & Topping, K. (1998). Peers assessing peers: Possibilities and problems. In K. Topping & S. Ehly (Eds.), *Peer-assisted learning.* Mahwah, NJ: Lawrence Erlbaum.

O'Malley, J. M., & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: Practical*

approaches for teachers: New York: Addison-Wesley.

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

Stanley, J. (1992). Coaching student writers to be effective peer evaluators. *Journal of Second Language Writing, 1*(3), 217-233.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276.

Wright, B., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transaction of the Rasch Measurement SIG, 8,* 370.

Yamashiro, A. (1999). *Using structural equation modeling to validate a rating scale*. Paper presented at the The 21st Language Testing Research Colloquium, Tsukuba, Japan.

Appendix

Assessment of Individual Presentation (adapted from Yamashiro 1999)

	Excellent (4)	Good (3)	Average (2)	Need Work(1)
1) Pace(ペース) 丁度良いペースで進んだか	良いペースで進んだ	ペースが少し速い・遅い	ペースが やや速い・遅い	ペースがかなり速い・遅い
2) Intonation(抑揚) 適切な抑揚，休止だったか	適切な抑揚である・不自然な休止がない	だいたい適切な抑揚である・不自然な休止はすくない	やや適切でない抑揚がある・不自然な休止がある	不適切な抑揚が多くみられる・不自然な休止が多い
3) Diction(発音の明瞭さ) 明瞭な発音であったか	全体に明瞭な発音である	だいたい全体に明瞭な発音である	あまり明瞭でない発音がみられる	かなり明瞭でない発音がみられる
4) Posture(姿勢) 良い姿勢だったか	全体に良い姿勢である	だいたい良い姿勢である	あまり良い姿勢でない	かなり姿勢がよくない
5) Eye contact(視線) 聴衆一人一人の目を見ていたか	全体に聴衆をしっかり見ている	だいたい聴衆を見ている	あまり聴衆をみていない	ほとんど聴衆をみていない
6) Gesture(ジェスチャー) 適所にジェスチャーを用いたか	適所にジェスチャーが用いている	だいたい適所にジェスチャーが用いている	あまりジェスチャーがもちいられない	ほとんどジェスチャーがない
7) Introductional(導入) トピックと全体の概要が適切に提示されたか	トピックと全体の概要が適切に提示されている	トピックと全体の概要がだいたい適切に提示されている	あまりトピックと全体の概要が適切に提示されていない	ほとんどトピックと全体の概要が提示されていない
8) Body(内容) 内容が詳しく具体的に展開されたか	内容が詳しく具体的に展開されている	だいたい内容が詳しく具体的に展開されている	あまり内容が詳しく具体的に展開されていない	ほとんど内容が詳しく具体的に展開されていない
9) Conclusion(結論) 全体のまとめと結論が適切に提示されたか	全体のまとめと結論が適切に提示されている	だいたい全体のまとめと結論が適切に提示されている	あまり全体のまとめと結論が適切に提示されていない	ほとんど全体のまとめと結論が適切に提示されいない
10) Language Use(正しい文の使用) 聞き手にとって分かりやすい正しい文が用いられたか	聞き手にとって分かりやすい正しい文が用いられている	だいたい聞き手にとって分かりやすい正しい文が用いられている	あまり聞き手にとって分かりやすい正しい文が用いられていない	ほとんど聞き手にとって分かりやすい正しい文が用いられていない
11) Vocabulary(語彙) 聞き手にとって分かりやすい正しい語が用いられたか	聞き手にとって分かりやすい正しい語が用いられている	だいたい聞き手にとって分かりやすい正しい語が用いられている	あまり聞き手にとって分かりやすい正しい語が用いられていない	ほとんど聞き手にとって分かりやすい正しい語が用いられていない
12) Purpose(目的) この発表の目的を達成したか	この発表の目的を達成している	だいたいこの発表の目的を達成した	あまりこの発表の目的を達成していない	ほとんどこの発表の目的を達成していない
13) Visual aids(視覚的な補助教材) 絵，写真，模物を効果的に用いたか	絵，写真，模物を効果的に用いている	だいたい絵，写真，模物を効果的に用いている	あまり絵，写真，模物を効果的に用いていない	ほとんど絵，写真，模物を効果的に用いていない

[Abstract]

Rater Training Effects on Peer Assessment of EFL Individual Presentations : An Interim Report

Hidetoshi SAITO

This study investigated the effects of rater training on peer assessment of EFL presentations by addressing the following three research questions: (1) Is the instructors' rating more similar to the treatment group's peer rating than to the control group's? (2) Is the treatment group's rating more severe than the control group's because of training? (3) Is the treatment group's attitudes toward peer assessment more positive than those of the control group's? Seventy-eight freshmen university students were randomly divided into two groups. The treatment group received a forty-minute rater training session, while the control group worked on an irrelevant writing task. All the students and three instructors assessed individual presentations. The results indicated that there was no effect due to rater training, although the ratings of both control and treatment groups demonstrated fairly high convergence with teacher ratings and positive attitudes toward peer assessment.