Chapter 7
Learning by Assessing in an EFL Writing Class

Trevor A. Holster, William R. Pellowe, J. Lake, and Aaron Hahn

Abstract This study used many-faceted Rasch measurement to research peer assessment in EFL writing classes, following previous research which reported acceptance of the pedagogical benefits by students of attention paid to a rubric during peer assessment. Pre and post treatment writing was compared on two rubrics, one targeting specific instructional items, the second intended to measure general academic writing. Students used the instructional rubric to conduct peer assessment, but were not exposed to the secondary rubric. Substantively and statistically significant gains were observed on the instructional rubric but not on the secondary rubric, providing evidence of learning by assessing. Response patterns suggested holistic rating by peer raters, resulting in effective rank ordering of overall performances but an inability to provide formative feedback, supporting the view that the mechanism of learning was awareness arising from learning by assessing.

Keywords Peer assessment • Rasch model • Attention • Academic writing

7.1 Introduction

Standardized proficiency tests such as TOEFL (ETS 2008) aim to incorporate characteristics of language sampled from relevant real world tasks into test items (Chapelle 2008). Such tests focus on stable traits using multiple choice

T.A. Holster (✉) • J. Lake
Academic English Program, Fukuoka Women's University, Fukuoka, Japan
e-mail: trevholster@gmail.com

W.R. Pellowe
Department of Human Engineering, Kinki University, Higashiosaka, Japan

A. Hahn
Department of English, Fukuoka High School, Fukuoka, Japan

Q. Zhang and H. Yang (eds.), *Pacific Rim Objective Measurement Symposium (PROMS) 2012 Conference Proceeding*, DOI 10.1007/978-3-642-37592-7_7,

tests, but many testing experts also support performance tests, integrating multiple skills into a single performance (see Brown et al. 2002; Hughes 2003; McNamara 1996, for example). Essay writing, for instance, requires integration of vocabulary, grammar, rhetorical structure, and content knowledge into a single performance. While performance tests can provide summative scores, they are also asserted to have formative benefits on study or teaching (Brown and Hudson 2002; Hughes 2003), but this requires that teachers and students become aware of weaknesses in their performances and are provided with opportunities to remedy them. Thus, validity arguments about formative tests require demonstration that the results are of sufficient quality to guide learning or instruction. One advantage of Rasch models (Bond and Fox 2007) over classical test theory (CTT) is the provision of fit statistics showing the consistency of students, items, and raters relative to the overall dataset, as demonstrated by Engelhard's (2009) investigation of students with disabilities. In the case of judged performance tests, many-faceted Rasch measurement (MFRM) (Linacre 1994) therefore provides invaluable diagnostic information about students, raters, or items in need of remediation.

In addition to feedback from teacher ratings, performance assessment used in conjunction with peer assessment (PA) and self assessment has potential formative benefits. Topping describes "learning by assessing" (LBA) as forcing increased time on task "interrogating the product or output, evaluating it in relation to intelligent questions at a macro and micro level" (1998, p. 254), following Graesser et al.'s (1995) investigation of discourse patterns in peer tutoring. Yarrow and Topping (2001) in their study of collaborative writing found that children made improvements in writing using a paired writing system that incorporated metacognitive prompting and scaffolding through the use of a framework in the form of a flowchart. They suggested that the flowchart acts as a metacognitive prompt, thus is both a tool for "procedural facilitation" during the writing process and for "product-oriented instruction" (p. 267) during editing and evaluating stages as well as for peer and self assessment. Similarly, Crinon and Marin researched French primary school children's writing, finding that children assigned to give feedback made larger improvements, "achieving greater overall coherency", than those who just received feedback, who "frequently make very specific local edits" (2010, p. 117). Likewise, Li et al. (2010) found that giving quality peer feedback was related to improved projects while receiving quality feedback did not improve the quality of final projects.

Further evidence of LBA was reported by Y. Cho and K. Cho (2011) who investigated English native speaker undergraduates writing technical reports, noting that "the effects of received peer comments were limited" (p. 639), but that writing improved more by giving comments than receiving comments. Further to this, K. Cho and MacArthur (2010) reported greater improvement for students who received feedback from multiple peers than from a single peer or single expert. The relative ineffectiveness of expert feedback suggests that it is engagement in the assessment process that provides the mechanism of learning rather than the product of assessment. Furthermore, K. Cho and MacArthur (2011)

compared outcomes for students who reviewed versus those who merely read the writing of peers and a control group, finding statistically and substantively better writing outcomes from the reviewing group, results that "clearly support the learning-by-reviewing hypothesis" (p. 77). K. Cho and Schunn (2010, pp. 209–210) noted that:

> The less obvious shift involves the change from students being the *receivers* of instructional explanations to students being the *generators* of instructional explanations. . . . In the role of reviewer, a student engages in reading, text analysis, and writing. . . . Coming to understand the criteria well enough to apply them to another student's paper provides students with the opportunity to improve their own writing and revision activities.

PA has attracted steady interest in second language acquisition. Mendonça and Johnson (1994) found that students not only made revisions due to peer comments but also self-noticed problems during peer review negotiations. They found that peer interactions forced students to be more active in their thinking about writing and this leads them to be able to use their knowledge in their revisions. Diab and Balaa (2011) used rubrics as instructional tools and for formative self assessment and peer assessment in EFL writing classes in Lebanon, finding statistically significant improvements in second-draft scores and strong endorsement from students of the value of the rubrics as learning tools. However, Min, studying Taiwanese university students, noted that although there are a large number of studies showing the benefits of peer response, "few studies have examined the extent to which peer feedback is incorporated into students' subsequent revisions" (2006, p. 119). Yang et al. (2006) compared teacher and peer feedback among Chinese EFL writing students, finding that teacher feedback was more likely to be incorporated and led to greater improvement, but also that it led to more superficial revisions. Peer feedback was found to be more likely to lead to meaning-change revision, a result attributed to negotiation of meaning during the peer interaction. Importantly, students in the peer feedback group had a much more positive view of its usefulness than students in the teacher feedback group.

Although a number of studies provide support for the formative benefits of peer assessment, the performance of student raters leads to doubts about its use for summative assessment. Roskams (1999) found support for LBA, but less support for assigning summative grades, while Tsui and Ng (2000) found that students preferred teacher responses over peer responses and incorporated more teacher feedback into revisions. They also found that students benefited more from reading the writing of peers than from peers' written comments, consistent with LBA. Mok (2011) found some acceptance of LBA among junior high school students, but serious concerns overall about the implementation of PA, while Cheng and Warren (2005) found students to be uncertain of their ability to rate peers and a tendency to rate holistically. Wong Mei Ha and Storey (2006) used journals to encourage reflection on self and peer editing, comparison with their own writing, and reflection on changes. Metacognitive awareness increased through reflection linking their declarative knowledge to their procedural knowledge. Saito and Fujita (2004)

compared peer and teacher ratings, finding a moderate to strong correlation, some support for LBA, but greater confidence in teacher ratings. Saito (2008) investigated the effect of rater training among Japanese university students, finding peer assessment to be effective overall, but a small effect from rater training and misfit patterns suggesting differential rubric interpretation between teachers and peer raters. Fukuzawa (2010) found acceptable fit for high school peer raters, but patterns suggesting a reluctance to use the lower categories on the rating scale, while Hirai et al. (2011) found undergraduate peer raters to be more lenient than teachers, as well as evidence of differential rating. Further concern over the consistency of peer raters was raised by Farrokhi et al.'s (2012) study of Iranian university students' English compositions, with student raters showing "a pattern of severity and lenience toward items that is opposite to that of teachers" (p. 93), and suggested the possibility that "they did not have a clear understanding of the assessment criteria."

Thus, although there is evidence supporting the effectiveness of LBA, serious doubts remain about students' understanding of the rubric, raising the question of the underlying mechanism by which LBA might aid second language acquisition. Schmidt (1990) argued that "noticing" driven by conscious attention is necessary for acquisition. Consistent with this, Schuchert (2004) argued that attention has a neurobiological basis, requiring alignment of five elements: an overall behavioral goal, a task-related goal, motor planning, stimulus qualities, and assessment of the influences of the four previous elements. This alignment produces the noticing required for both initial and advanced learning, so attention must be maintained in multiple sessions over extended timeframes for new knowledge to be consolidated as procedural knowledge. Although student raters may interpret the rubric differently to teachers, LBA may promote alignment of the elements of attention, leading to improved awareness of the rubric and noticing of the difference between students' own performances and target language features.

Thus, rather than the emphasis on the measurement of ability and consistency of test items typical of summative test analysis (Bachman 1990; Henning 1987), LBA's success rests on the quality of interaction of assessors with the rubric. Inconsistent assessors may have misunderstood the rubric or employed it idiosyncratically, casting doubt on their feedback to peers or ability to benefit from LBA. A validity argument for an assessment intended to generate LBA must therefore demonstrate that this interaction results in formative benefits independently of the quality of summative measurement. MFRM analysis provides an elegant solution to this, allowing peer raters' performances to be compared against those of teachers to ascertain whether the same trait is assessed by the different groups of raters, while also providing interval level measurement. The logit outputs provided by Rasch analysis provide convenient measures of effect size, allowing comparison between different studies on the basis of substantive meaningfulness, addressing Thompson's (1999) critique of misuse of measures of statistical significance, which are highly dependent on sample size.

7.2 The Study

7.2.1 Research Questions

RQ1: Do student performances improve after PA using the instructional rubric?
RQ2: Is peer feedback on specific rubric items comparable to teacher feedback?
RQ3: Are gains in the instructional rubric reflected in writing proficiency overall?

7.2.2 Background and Method

This study was conducted in writing classes in an Academic English Program (AEP) at a public women's university in Japan, but no familiarity with paragraph length organization could be assumed prior to this course. The participants ($N = 30$) in this convenience sampling were assigned to second semester classes taught by one of the authors, comprising 45 hours of instruction over 15 weeks. Although writing classes were conducted in two class groups of 15 students, all 30 students concurrently took a listening class together. The course book, *Ready to Write 3* (Blanchard and Root 2010), included brief grammar reviews which were used in class, but intensive grammatical instruction was not attempted due to time constraints and concerns over the sequencing and teachability of grammar features in general. Instruction targeted organizational features of writing considered to be learnable through explicit attention. Classes therefore focused on reviewing paragraph and essay organization and providing extensive writing practice on topics related to students' everyday experiences. Quite general topics were assigned, and students were encouraged to incorporate their personal experiences in order to maintain the relevance, novelty and coping potential argued by Schumann and Wood (2004) to underpin long-term motivation, while providing the alignment of the elements of attention described by Schuchert (2004).

Following reviews and practice of paragraph organization, the students planned and wrote three body paragraphs on the topic of "planning a trip that is educational, economical, and enjoyable". These were then combined into a complete essay following explanations of introductory and concluding paragraphs. A supplementary workshop on formatting conventions and the use of word processing software was conducted separately from the textbook curriculum, followed by peer review of the complete essays and submission of revised drafts for the PA session in the next class. Twenty-six students submitted Essay 1 in time, so these were randomly distributed to students for PA using the rubric shown in the Appendix.

PA generates very large numbers of responses, so data was collected using the peer assessment module for the open source MOARS audience response system (Pellowe 2002, 2010a, b). This system can output data ready for immediate MFRM analysis, making MFRM analysis practical within minutes of students completing their performances (Holster and Pellowe 2011). Paper rating sheets were used in

conjunction with MOARS, allowing items to be rated non-sequentially and providing a backup in case of technical issues with the online system. After approximately an hour of PA, students accessed the ratings for their own essay, presented in the form of bar graphs, and were asked to note strong and weak areas of their performances. Students were not provided the teacher's ratings separately from the PA results.

Division-classification essays were reviewed next, and then students were assigned the topic "Bad Habits" for Essay 2, required to plan three supporting paragraphs in class, and assigned a first draft for homework. In the next class, an introduction and conclusion were added and students were given a further week to produce a final draft. Twenty-four essays were submitted by the deadline; 23 students completed both Essay 1 and Essay 2.

7.2.3 Analysis and Results

Reliance of performance tests on human raters raises issues of rater performance. McNamara (1996) and Weigle (1994) provided seminal accounts of the use of MFRM to monitor rater performance and adjust for differences in severity. While teachers and students are implicitly familiar with two-faceted tests, where the probability of a successful response results from the interaction of student ability and item difficulty, judged performances introduce a third facet of measurement. The resulting probability of success is modeled as:

$$P = \exp(B - D - R)/(1 + \exp(B - D - R))$$

where P represents the probability of success, B represents person ability, D represents item difficulty, and R represents rater severity (Linacre 1994). Consistent with intuition, the odds of success increase with person ability, but decrease with item difficulty or rater severity. For the current study, a fourth facet, "Time", was included, on the hypothesis that student ability would increase as a result of the PA following Essay 1, and thus the probability of success would increase for Essay 2. This can be modeled as:

$$P = \exp(B - D - R - T)/(1 + \exp(B - D - R - T))$$

Three sets of teacher ratings were used for the initial analysis. Although the classroom teacher, T1, rated each essay as it was received, this resulted in multiple rating sessions over a period of several weeks, reflecting classroom reality, but raising concern over the consistency of the subsequent rating performance. T1 therefore rated all the essays again in a random order 1 week after the submission deadline for Essay 2. These ratings are indicated by T1A and T1B for the first and second ratings, respectively. For comparison, a second AEP teacher, T2, also rated all the essays following final submission.

MFRM analysis allowed measures of student ability, rater severity, proficiency gains across time, and item difficulty to be measured on a common log odds, or "logit", scale, representing equal interval measures, with items centered on 0.00 logits. Engelhard (2009) suggested 0.30 logits as a threshold for a substantive effect size, and all four facets showed substantively meaningful ranges. T1A was 0.68 logits more lenient than T2, an effect size translating into relative probabilities of success of 59 % versus 41 % for an item of average difficulty. The range of item difficulty, 2.31 logits, was extremely large. A student having a 50 % expectation of success with "Conclusion" (1.02 logits) would have a 91 % expectation of success on "Formatting" (−1.29 logits). Even the least proficient student (−0.66 logits) was substantively more able than "Formatting" is difficult, so this item provides little information about the ability of this sample of students. Finally, a provisional answer to RQ1 is possible by looking at the facet of "Time": student performances improved by approximately 1 logit following the PA session, a substantively very large gain.

However, definitive answers to the research questions assume acceptable functioning of all facets, so more detailed investigation is warranted. Fundamental to MFRM are assumptions of a unidimensional trait, so a preliminary question is whether teachers' ratings meet this requirement, given that these provide the benchmark against which PA ratings are compared. Item fit statistics provide an indication of whether the rubric describes a unidimensional trait, while rater fit statistics allow analysis of rater performance. Of particular concern was rubric Item 10, "Formatting", which addressed the use of word processing software rather than language proficiency. Rasch item analysis of the rubric was therefore conducted to determine whether psychometric evidence supported the content based argument concerning the dimensionality of the rubric. Table 7.1 shows the measurement report for items, ordered by model-data fit, shown by the infit and outfit mean-square (*MS*) statistics. "Formatting" is the most misfitting item, with infit and outfit statistics of 1.58 and 1.41 respectively. Given the questionable content validity of this item, this level of misfit supports removal of this item from the analysis.

Rater performance was investigated next, shown in Table 7.2. Raters T2 and T1B are slightly more consistent than expected, with mean-square statistics below 1.00, but T1A, with respective values of 1.17 and 1.14, is slightly misfitting. While this does not threaten overall measurement, the variation in performance between T1A and T1B shows the importance of multiple ratings for performance assessments. Subsequent analyses use only the ratings from T1B and T2.

Having demonstrated acceptable data-model fit for teacher ratings, the PA data was analyzed next. As peer feedback derived only from ratings of Essay 1, student ability measures for this essay were compared for teacher ratings and peer ratings, plotted in Fig. 7.1. Considerable agreement in rank ordering between teacher ratings and peer ratings is apparent, with a raw correlation of .87 indicating 75 % shared variance between the two sets of measures. With reliability coefficients for person measurement of .91 for peer ratings and .89 for teacher ratings, the disattenuated correlation rises to .97, indicating effectively interchangeable rank

Table 7.1 Item measurement report from ratings by teachers

Items		Score	*n*	*M*	Logit measure	*SE*	Infit *MS*	Outfit *MS*	Pt-meas corr
10	Formatting	355	149	2.4	−1.29	0.14	1.58	1.41	.44
2	Introduction	224	150	1.5	0.73	0.12	1.21	1.20	.46
3	Conclusion	203	150	1.4	1.02	0.12	1.11	1.10	.40
1	Thesis stment	246	150	1.6	0.43	0.12	1.10	1.09	.57
6	Support	252	150	1.7	0.35	0.12	0.98	1.00	.26
7	Coherence	282	150	1.9	−0.08	0.12	0.90	0.90	.48
8	Cohesion	286	150	1.9	−0.14	0.12	0.89	0.89	.46
5	Unity	299	150	2.0	−0.33	0.12	0.83	0.82	.60
9	Relevance	269	150	1.8	0.11	0.12	0.80	0.80	.48
4	Organization	329	150	2.2	−0.80	0.13	0.73	0.74	.59
M ($n = 10$)		274.5	149.9	1.8	0.00	0.12	1.01	1.00	.47
SD (Pop)		43.9	0.3	0.3	0.66	0.01	0.24	0.20	.10
SD (Sample)		46.2	0.3	0.3	0.69	0.01	0.25	0.21	.10

Model(Pop): RMSE .12 Adj (True) SD.65 Separation 5.30 Strata 7.40 Reliability .97
Model(Samp): RMSE .12 Adj (True) SD.68 Separation 5.60 Strata 7.80 Reliability .97
Model fixed (all same) chi-square: 271.7 *df*: 9 significance (probability): .00

Table 7.2 Teacher rater's measurement report

Raters	Score	*n*	*M*	Logit meas	*SE*	Infit *MS*	Outfit *MS*	Pt-meas corr
T1A	881	450	2.0	−0.43	0.07	1.17	1.14	.59
T2	773	450	1.7	0.12	0.07	0.94	0.94	.43
T1B	736	450	1.6	0.31	0.07	0.89	0.90	.57
M ($n = 3$)	796.7	450.0	1.8	0.00	0.07	1.00	0.99	.53
SD (Pop)	61.5	0.0	0.1	0.31	0.00	0.12	0.11	.07
SD (Sample)	75.3	0.0	0.2	0.38	0.00	0.15	0.13	.09

Model(Pop): RMSE .07 Adj (True) SD .31 Separation 4.29 Strata 6.05 Reliability .95
Model(Samp): RMSE .07 Adj (True) SD .38 Separation 5.30 Strata 7.40 Reliability .97
Model fixed (all same) chi-square: 57.2 df: 2 significance (probability): .00
Inter-Rater Exact Agreements: 595 = 44.1% Expected: 552.4 = 40.9%

ordering within the limits of measurement error. However, it is also apparent from Fig. 7.1 that peer raters were much more lenient than teachers, with mean logit measures of 1.79 for peer ratings versus −0.04 for teachers, corresponding to mean raw ratings of 2.3 versus 1.5 on the rating scale of 0–3.

In contrast, although the teachers and peer raters returned very similar rank ordering of person measures, this did not hold for the ranking of item difficulty measures, as shown in Fig. 7.2. While the range of item difficulty from teacher ratings was 2.12 logits, the range from peer ratings was only 0.83 logits, raising doubts about peer raters' interpretation of the rubric. The respective mean rating for

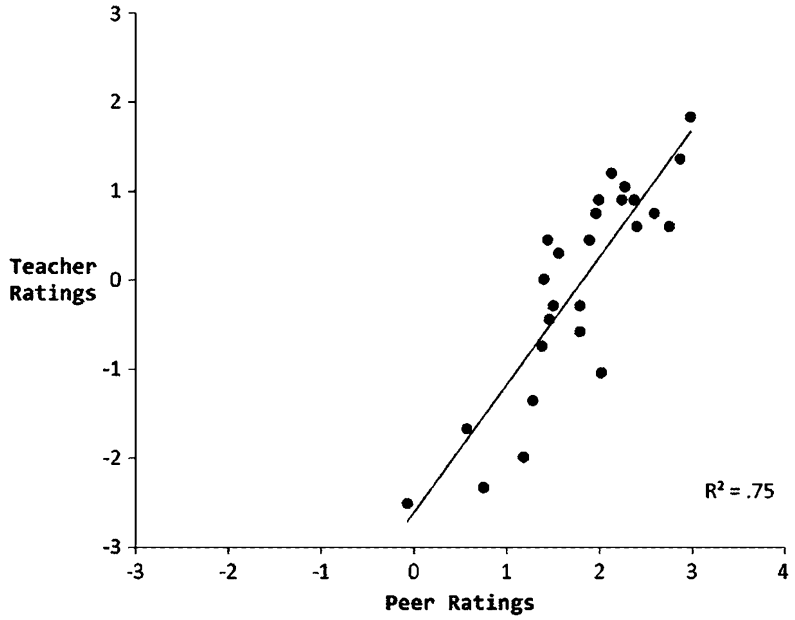

Fig. 7.1 Person ability measures from teacher ratings and peer ratings. Ratings of writing ability made by teachers and peers are mapped, showing a strong linear trend-line with shared variance of 75 %

teachers and peer raters were 1.5 and 2.4, with standard deviations of 0.2 and 0.1, and ranges of 0.8 and 0.3. Peer raters avoided the lower categories on the rating scale, while teachers were more likely to utilize the full range of the scale. These results are consistent with holistic ratings by peers, resulting in discrimination of good performances from poor, but not between the items on the rubric. Thus, RQ2 is answered: peer feedback on specific rubric items is not comparable to teacher feedback. Given that teacher feedback was not provided in this case, the inability of peer raters to provide diagnostic feedback raises the question of the source of the large gain between Essay 1 and Essay 2.

Figure 7.3 plots the interaction between items and time from teacher ratings, with all rubric items receiving higher mean ratings for Essay 2, reflected in the lower measures of difficulty. This further confirms RQ1: student performances following PA improved substantively. Only one item, "Thesis stment", with a gain of 2.18 logits, showed substantively larger improvement than the mean of 1.04 logits, but comparison with Fig. 7.2 shows that peer raters rated this as relatively easy, unlike teachers who rated it as difficult. Peer feedback cannot therefore have signaled to students that this item needed remediation, evidence against peer feedback as a major mechanism of improvement.

Given that Essay 1 was many students' first attempt at writing essay length compositions, the question arises whether the substantive overall gains arose from practice rather than LBA. Therefore a secondary rubric was developed independently to measure general writing proficiency. A writing instructor, R1, with a Masters degree in writing instruction and experience teaching academic writing to both North American university undergraduates and L2 learners in Japan was shown the submissions for

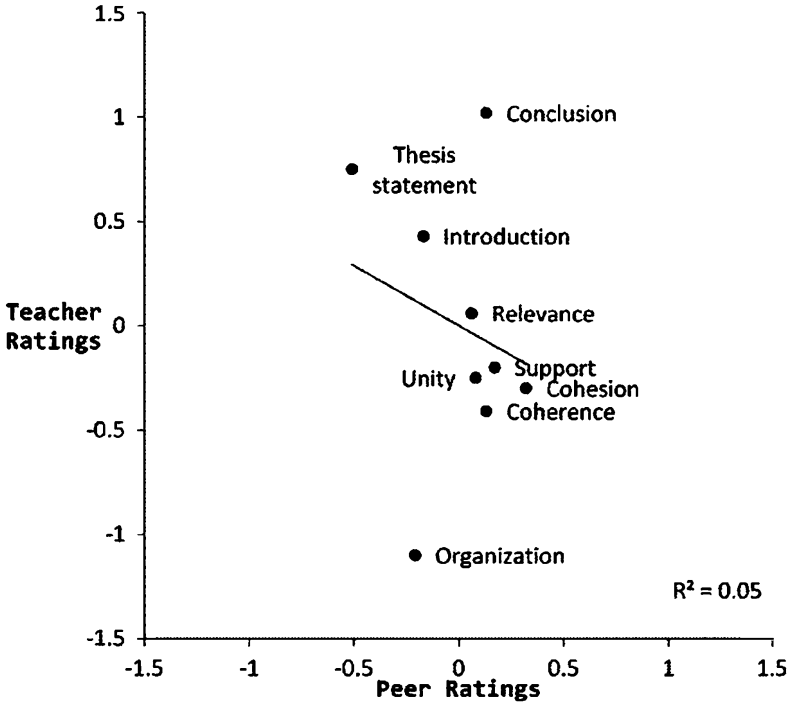

Fig. 7.2 Item difficulty measures from teacher ratings and peer ratings. The difficulty of rubric items estimated from teacher ratings and peer ratings are mapped, showing no correlation between the two sets of measures

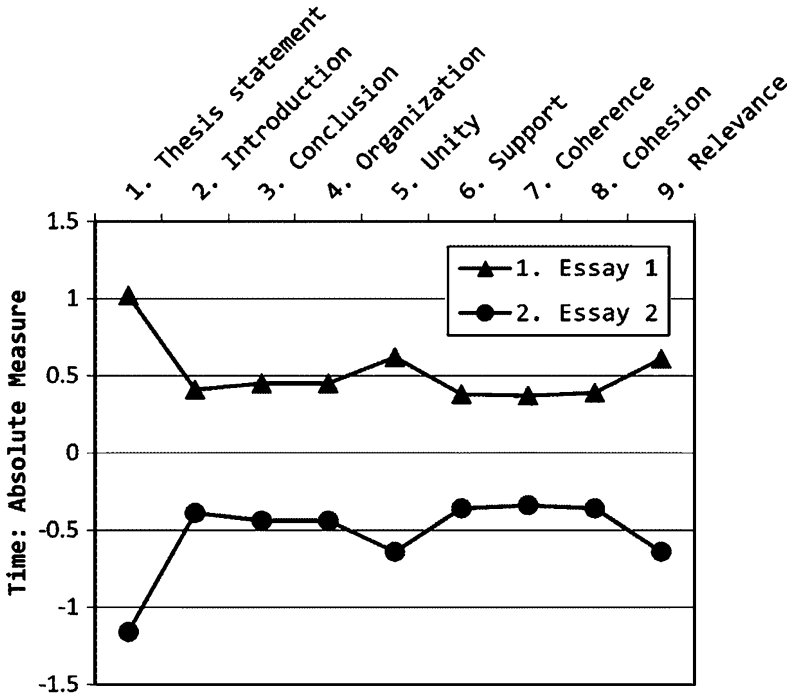

Fig. 7.3 Change in item difficulty by time for instructional rubric. The difficulty of rubric items is compared for Essay 1 and Essay 2. Ratings were substantively higher on all items in Essay 2, evidence of learning related to the rubric items

Table 7.3 Secondary rubric raters' measurement report

Raters	Score	*n*	*M*	Logit meas	*SE*	Infit *MS*	Outfit *MS*	Pt-meas corr
R1	670	195	3.4	0.01	0.07	1.16	1.16	.62
R2	691	200	3.5	−0.01	0.07	0.83	0.83	.40
M ($n = 2$)	680.5	197.5	3.4	0.00	0.07	0.99	0.99	.51
SD (Pop)	10.5	2.5	0.0	0.01	0.00	0.17	0.17	.11
SD (Sample)	14.8	3.5	0.0	0.02	0.00	0.24	0.24	.15

Model (Pop): RMSE .07 Adj (True) SD .00 Separation .00 Strata .33 Reliability .00
Model (Samp): RMSE .07 Adj (True) SD .00 Separation .00 Strata .33 Reliability .00
Model fixed (all same) chi-square: .1 *df*: 1 significance (probability): .82
Inter-rater agreement opportunities: 195 Exact agreements: 52 = 26.7 % Exp: 53.8 = 27.6 %

Essay 1, but told only that students were taking academic writing classes and that they were asked to write an essay on the topic of "Planning a Trip", with introductory and concluding paragraphs and a minimum of three supporting paragraphs. The resulting rubric was based on experience with the writing section of the GRE (ETS 2012) and comprised four operational items, "Grammar", "Organization/Structure/Length", "Vocabulary/Register/Tone", and "Content/Logic/Context", rated on a scale from 1 to 6. A second rater, R2, was used to provide inter-rater comparison, this rater having an undergraduate degree in literature and a Masters degree in applied linguistics, with over two decades experience teaching L2 learners of English in North America and Japan.

Table 7.3 shows the rater measurement report for the secondary ratings. An inconsequential difference of 0.02 logits was found in severity. R1 was slightly less consistent than expected, with infit and outfit mean square statistics of 1.16 indicating 16 % more randomness than modeled, while R2 was correspondingly overfitting, with infit and outfit statistics of 0.83, levels not threatening to effective measurement. Table 7.4 shows the measurement report for items. The mean of both infit and outfit was 0.99, both having extremely low standard deviations of 0.09, while the most misfitting item was "Grammar", with infit and outfit mean square statistics of 1.13. Reliability of person measurement was .78, so these items are functioning acceptably and able to separate low ability persons from high.

Analysis of the secondary ratings found Essay 1 to be 0.08 logits more difficult than Essay 2, a gain that was neither statistically nor substantively significant. The gains measured in the PA rubric were not replicated, consistent with gains resulting from explicit awareness of the rubric rather than general proficiency. RQ3 is thus answered: gains in the instructional rubric were not replicated in the secondary rubric. However, Fig. 7.4 shows the time versus item interaction, with "Grammar" given lower ratings and "Organization/Structure/Length" given higher ratings in Essay 2. Drawing firm conclusions about this from such a small pilot dataset is inadvisable, but it is notable that "Organization/Structure/Length" is similar to PA rubric items, while the others are not, consistent with PA leading to LBA.

Table 7.4 Secondary rubric items' measurement report

Items	Score	*n*	*M*	Logit meas	*SE*	Infit *MS*	Outfit *MS*	Pt-meas corr
2 Grammar	339	99	3.4	0.02	0.10	1.13	1.13	.51
3 Org/struc/length	354	98	3.6	−0.16	0.10	0.99	0.98	.51
4 Vocab/reg/tone	359	99	3.6	−0.17	0.10	0.88	0.90	.47
5 Cont/logic/context	309	99	3.1	0.32	0.10	0.96	0.96	.51
M ($n = 4$)	340.3	98.8	3.4	0.00	0.10	0.99	0.99	.50
SD (Population)	19.5	0.4	0.2	0.20	0.00	0.09	0.09	.02
SD (Sample)	22.5	0.5	0.2	0.23	0.00	0.10	0.10	.02

Model (Pop): RMSE .10 Adj (True)SD .17 Separation 1.72 Strata 2.63 Reliability .75
Model (Samp): RMSE .10 Adj (True)SD .21 Separation 2.07 Strata 3.09 Reliability .81
Model, Fixed (all same) chi-square: 15.8 df: 3 significance (probability): .00

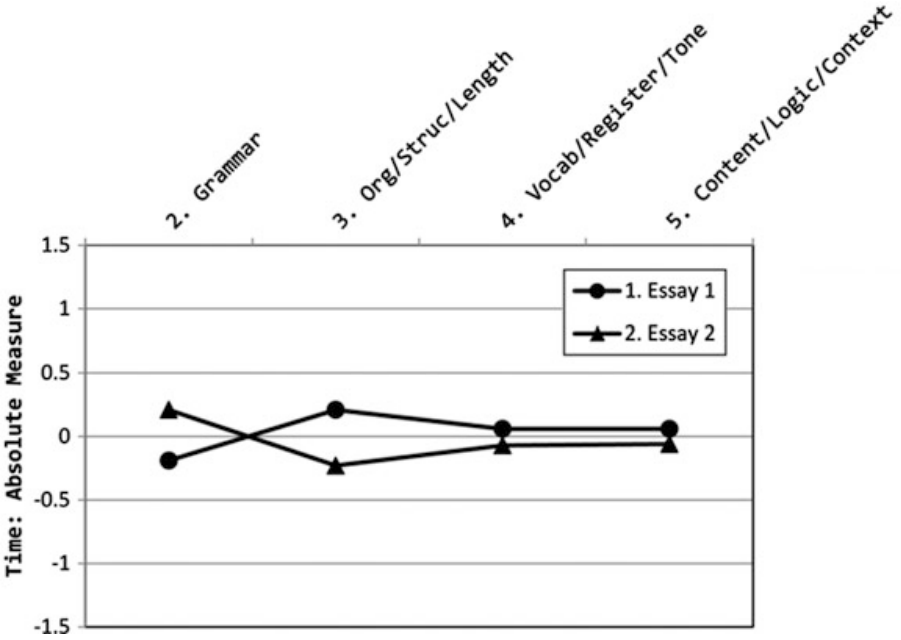

Fig. 7.4 Change in item difficulty by time for secondary rubric. The difficulty of rubric items is compared for Essay 1 and Essay 2, with a substantive gain on Item 3 offset by a substantive loss on Item 2

7.3 Discussion and Implications

The major research question, RQ1, concerned the effectiveness of PA leading to improved performance on Essay 2. The results supported this, a substantively large effect size occurring between Essay 1 and Essay 2 on the PA rubric but not on the secondary rubric, consistent with LBA. As teacher feedback was not provided on

the rubric items and teachers and peer raters employed the rubric differently, peer feedback could not have identified rubric items in need of remediation, leaving attention to the rubric during the rating sessions as the most plausible source of learning. Although the research design did not control for the difficulty of the essay topics, raising the concern that the higher ratings for Essay 2 on the PA rubric may have resulted from the second topic being easier than the first, this pattern was not seen in the results from the secondary rubric. This supports LBA as a powerful mechanism of learning through drawing attention to key features of performances.

These results support the validity of peer assessment as a classroom instructional task while holding potential benefits for motivation because it was interaction with samples of student language that resulted in LBA. This expands the input available to learners and addresses the argument for a balance between familiarity and novelty made by Schumann and Wood (2004) by providing input on topics that are relevant and interesting while promoting the alignment of the elements of attention described by Schuchert (2004).

However, the results of this pilot study were limited by sampling constraints and the limited timeframe. The classroom teacher's impressionistic feeling was that these students had very high intrinsic motivation and were not representative of average Japanese university students. Although large gains were observed on the PA rubric between Essay 1 and Essay 2, a ceiling effect may occur if further rounds of writing and PA were administered, while it's plausible that larger gains on the secondary rubric would be observed over a longer timeframe. Furthermore, the vagueness of the essay prompts, intended to provide students with opportunities to write about familiar topics and experiences, were not well suited to the secondary rubric, based on the expectations of L1 academic writing. Addressing these issues was beyond the scope of this pilot study, but highlight the need for a larger scale quasi-experimental study to confirm these findings and provide evidence of wider generalizability.

Appendix

Essay Rating Instructions and Rubric

Essay Revision

Read other students' essays. Rate each essay from "A" to "D" on the following points by marking the bubbles on the grading sheet.

他の学生の発表を見て評価をします。以下の評価基準を参考にして、評価シートのA~Dをりつぶして下さい。

"A" = Excellent performance.

(素晴らしい。)

"B" = Good performance, but could be improved.

(良いが、改善出来る部分もある。)

"C" = Weak performance, should be improved.

(良いとは言えない。改善した方が良い。)
"D" = Very weak performance, must be improved.
(良くない。改善すべき。)

1. **Thesis stment**: How well does the introduction identify the focus of the essay using a thesis stment?
2. **Introduction**: How well does the introduction preview the main points of the essay?
3. **Conclusion**: How well does the conclusion summarize the main points of the essay?
4. **Organization**: Are the supporting paragraphs in a logical order?
5. **Unity**: Does each supporting paragraph have a clear topic sentence and focus?
6. **Support**: Do the supporting paragraphs support the essay focus with specific details?
7. **Coherence**: Are the supporting sentences in each paragraph organized in a logical way?
8. **Cohesion**: Did the writer use transition words to guide the reader from one idea to the next?
9. **Relevance**: Are all the supporting sentences relevant?
10. **Formatting**: Is the essay formatted correctly?

References

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Blanchard, K., & Root, C. (2010). *Ready to write 3: From paragraph to essay* (3rd ed.). White Plains: Pearson Longman.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). London: Lawrence Erlbaum Associates.

Brown, J. D., & Hudson, T. D. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. J. (2002). *An investigation of second language task-based performance assessments*. Honolulu: University of Hawaii.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). New York: Routledge.

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

Cho, Y., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39* (5), 629–643. doi:10.1007/s11251-010-9146-1.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328–338.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103* (1), 73–84. doi:10.1037/a0021950.

Cho, K., & Schunn, C. D. (2010). Developing writing skills through students giving instructional explanations. In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines*. New York: Springer.

Crinon, J., & Marin, B. (2010). The role of peer feedback in learning to write explanatory texts: Why the tutors learn the most. *Language Awareness, 19*(2), 111–128. doi:10.1080/09658411003746604.

Diab, R., & Balaa, L. (2011). Developing detailed rubrics for assessing critique writing: Impact on EFL university students' performance and attitudes. *TESOL Journal, 2*(1), 52–72. doi:10.5054/tj.2011.244132.

Engelhard, G. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585–602. doi:10.1177/0013164408323240.

ETS. (2008). The TOEFL® Test – Test of English as a Foreign Language™. Retrieved from http://tinyurl.com/zocgc. Accessed 28 Mar 2008.

ETS. (2012). About the GRE® revised General Test. Retrieved from http://www.ets.org/gre/revised_general/about. Accessed 19 Jan 2012.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal, 34*(1), 79–101.

Fukuzawa, M. (2010). Validity of peer assessment of speech performance. *Annual Review of English Language Education in Japan, 21*, 181–190.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495–522. doi:10.1002/acp.2350090604.

Henning, G. (1987). *A guide to language testing*. Boston: Heinle & Heinle.

Hirai, A., Ito, N., & O'ki, T. (2011). Applicability of peer assessment for classroom oral performance. *JLTA Journal, 14*, 41–59.

Holster, T. A., & Pellowe, W. R. (2011). *Using a mobile audience response system for classroom peer assessment*. Paper presented at the JALT CALL 2011 conference, Kurume University, Kurume.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology, 41*(3), 525–536. doi:10.1111/j.1467-8535.2009.00968.x.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Pearson Education.

Mendonça, C. O., & Johnson, K. E. (1994). Peer review negotiations: Revision activities in ESL writing instruction. *TESOL Quarterly, 28*(4), 745–769.

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15*(2), 118–141. doi:10.1016/j.jslw.2006.01.003.

Mok, J. (2011). A case study of students' perceptions of peer assessment in Hong Kong. *ELT Journal, 65*(3), 230–239. doi:10.1093/elt/ccq062.

Pellowe, W. R. (2002). *Keitai-assisted language learning (KALL)*. Paper presented at the 28th JALT international conference, Granship Conference Center, Shizuoka.

Pellowe, W. R. (2010a). MOARS (Version 0.8.3) [Audience response system]. Retrieved from http://moars.com

Pellowe, W. R. (2010b). *Quiz and survey system for mobile devices*. Paper presented at the 36th JALT international conference, WINC, Aichi.

Roskams, T. (1999). Chinese EFL students' attitudes to peer feedback and peer assessment in an extended pairwork setting. *RELC Journal, 30*(1), 79–123. doi:10.1177/003368829903000105.

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

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31–54. doi:10.1191/1362168804lr133oa.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. doi:10.1093/applin/11.2.129.

Schuchert, S. A. (2004). The neurobiology of attention. In J. H. Schumann, S. E. Crowell, N. E. Jones, N. Lee, S. A. Schuchert, & L. A. Wood (Eds.), *The neurobiology of learning* (pp. 143–174). Mahway: Lawrence Erlbaum Associates.

Schumann, J. H., & Wood, L. A. (2004). The neurobiology of motivation. In J. H. Schumann, S. E. Crowell, N. E. Jones, N. Lee, S. A. Schuchert, & L. A. Wood (Eds.), *The neurobiology of learning* (pp. 23–42). London: Lawrence Erlbaum Associates.

Thompson, B. (1999). Statistical significance tests, effect size reporting and the vain pursuit of pseudo-objectivity. *Theory & Psychology, 9*(2), 191–196. doi:10.1177/095935439992007.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276. doi:10.3102/00346543068003249.

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*(2), 147–170.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197–223. doi:10.1177/026553229401100206.

Wong Mei Ha, H., & Storey, P. (2006). Knowing and doing in the ESL writing class. *Language Awareness, 15*(4), 283–300.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200.

Yarrow, F., & Topping, K. J. (2001). Collaborative writing: The effects of metacognitive prompting and structured peer interaction. *The British Journal of Educational Psychology, 71*, 261–282.