An investigation into the use of Rasch analysis to aid L2 writers in anonymized peer-assisted learning

Jeffrey Martin
jeffmjp@gmail.com
J. F. Oberlin University

Abstract

This study critically evaluated an anonymized peer feedback and assessment design for L2 writers enhanced by the use of Rasch analysis. This approach centered on the acts of giving assessment (Topping, 1998) and feedback (Lundstrom & Baker, 2009) in exchange with multiple peers. Each participant received feedback comments and class-wide statistical measures summarized for students without the need for their peers to rate all papers. Anonymity was maintained to bring unencumbered attention to the role of the reader (Booth et al., 2008) and to provide a space for interpretation and reflection on the potentially contrasting data and experiences that emerge. This process is argued to drive cognitive development and improve L2 writing skills. An initial trial with 15 high-proficiency EFL learners indicated that the design facilitated an effective exchange for each participant. The effects of anonymously including teacher comments also brought informative insights about the perception of feedback and its sources. Issues were found regarding overly narrow use of the ratings scales by some participants. A 6-point rating scale is proposed for more differentiated scoring. Overall, positive engagement and reception by the participants suggests that this peer assisted learning approach holds promise for L2 writers.

Keywords: peer feedback, peer assessment, L2 writing, anonymous feedback/assessment, Rasch analysis, judging plan

Peer feedback and peer assessment activities invite L2 student writers to develop their ability to better self-evaluate their work and practice the skill of discernment when reading the work of others. Giving feedback (Lundstrom & Baker, 2009) and *learning by assessing* (Topping, 1998, p. 254) may be critical drivers of these developments. Additionally, Booth et al. (2008) argued that effective writers embody the role of their "distant readers" in the world (p. 280). Recognizing the variety of possible reader responses, this distance can be simulated in a learning environment through varied, anonymous exchange. Through the giving and receiving of comments and ratings by classroom peers, potentially contrasting information becomes available to the L2 writing student, an opportunity not always forthcoming in real-world communication. Fundamentally, the student-centered and formative benefits of peer feedback and peer assessment activities have been theorized and empirically tested in SLA research (Hyland & Hyland, 2006; Liu & Hansen, 2002; Yu & Lee, 2016). While there are distinctions between feedback and assessment (Falchikov, 2001), these approaches are viewed here as complimentary. To achieve the different aims proposed by Topping (1998) and Booth et al. (2008), the current study investigated the potential of incorporating anonymity and Rasch analysis within a judging plan (Linacre, 1989; Rasch, 1961). These were used to capture the characteristics described above and place them within a peer feedback and assessment process that can be feasible for both L2 student writers and instructors.

Rasch analysis is a statistical tool that provides class-wide and individual measures calibrated for both writing ability and rating behavior, all of which can be provided back to students anonymously together with qualitative feedback. The output from Rasch analysis is achieved with relatively limited rater data from each student and the instructor prepares this using Rasch analysis software. To further emphasize the student-centered environment, teacher feedback and ratings can be intermixed anonymously within the data. Such conditions provide opportunities for L2 learners to engage with the levels of Bloom's taxonomy (Bloom, 1956; Krathwohl, 2002) in a number of ways as they make judgements at several stages along the peer-assisted learning process. Importantly, outcomes of the Rasch analysis invite students to individually revisit the papers they previously reviewed and self-assess their judgements. This

paper critically evaluated an application of this approach in a semester-long academic course on business communication and leadership with 15 high-proficiency learners of English.

Literature review

Peer feedback and peer assessment

Peer feedback and peer assessment are associated terms that are viewed differently by researchers working under distinct theoretical frameworks. Research on peer-exchange often cites Vygotsky's zone of proximal development (ZPD, 1978) while other influential work builds on the cognitive developmental theory of Piaget (1971). The differences between these theories underpin the separation of peer feedback and peer assessment as applied by some in the field. This paper maintains that these activities are complementary and that the distinction between "giver" and "receiver" in peer-exchange may offer informative insights for both.

Differing views on peer feedback and peer assessment stem from different views on the learning process (Falchikov, 2001). Liu and Carless (2006) asserted that "peer feedback is primarily about rich detailed comments but without formal grades, whilst peer assessment denotes grading (irrespective of whether comments are also included)" (p. 280). Liu and Hansen (2002) outlined a Vygotskian perspective on peer feedback by defining it as a form of *peer response* where students take on reciprocal "roles and responsibilities normally taken on by a formally trained teacher, tutor, or editor" (p. 1). The resulting social interaction is what builds an individual's "current competence through the guidance of a more experienced individual" (p. 5). In this view, L2 student writers provide each other feedback and form a mutual ZPD within a sociocultural framework (Guerrero & Villamil, 1994). Student peers become "individually novices and collectively experts" for each other (Donato, 1994, p. 46), which allows for collective scaffolding that benefits the individuals within the group.

Topping and Ehly (1998) placed peer feedback and peer assessment together within a wider paradigm of *peer-assisted learning*, with Topping (1998) defining peer assessment as an arrangement for peers of equal status who consider the level, value, or worth of each other's products or outcomes of learning. In this view, peer-assisted learning finds grounding in the theories of Vygotsky (1978) but centers more on cognitive developmental theory (Piaget, 1971). The emphasis of the guiding expert(s) within a ZPD is lessened in Topping's paradigm. Rather, Topping argues that each learner is invited to reconcile differences between prior and new experiences through interactions with peers of relatively equal ability but with varied competencies. Peer assessment together with peer feedback can facilitate this symmetrical and reciprocal process where students may learn as much or more through the experience of engaging in the work of others than from the receiving of guidance. Foot and Howe (1998, p. 33) highlighted the Piagetian model of *cognitive conflict* as aptly describing this mechanism for peer-assisted learning. Topping (1998) additionally noted that the process emphasizes assessment to be a formative way to "maximize success rather than merely determine success or failure only after the event" (p. 249). In this Piagetian sense (1971), the learner enters a process of adaptation by reciprocally engaging with peers and their work. The cognitive conflicts that likely occur are points for the learner's cognitive development.

Anonymity and the role of the reader

As Zamel (1982) emphasized, the process of peer-exchange can develop in students "the crucial ability of re-viewing their writing with the eyes of another" (p. 206). Yu and Lee's (2016) overview of many studies, mostly featuring openly paired dyads, outlined how peer feedback activities benefit writing development. However, the social distractions of open peer-review can complicate a learner's perspectives on the sources of feedback. Intermixing peer feedback and teacher feedback, Xu and Liu

(2010) found that students took up anonymized feedback from both peers and teachers at similar rates. This is in contrast to the preference of some students who only value teacher feedback. Studies have also shown that anonymity can further influence the quality of peer-exchange. Rotsaert et al. (2018) found that the quality of peer feedback improved when it was given anonymously, suggesting that familiarity with the writer can inhibit genuine responses by peers. Affective factors were also found by Cheng and Tsai (2012) by showing how learners preferred anonymity to avoid social pressures. Additionally, it was essentially an anonymized peer review methodology in Lundstrom and Baker (2009) that demonstrated that the giving of feedback is potentially more effective for improving student writing than receiving it.

An important aspect for effective writing is the student's conceptualization of unknown readers outside of a classroom setting. This holds true for any writing domain. For example, Booth et al. (2008) emphasized that researchers and professionals write to achieve objectives with their readers, who are distant in the sense that they are often not known personally to the writer. Likewise in everyday business writing, Garner (2012) argued that writers more likely succeed by anticipating the "goals and priorities" of their readers (p. 7). This suggests that objectives and strategies vary widely by the style, purpose, and domain of writing, but the importance of the reader is constant. Logistically, the anonymous exchange in a writing course can be facilitated via digital file sharing techniques. Such exchange allows for data collection from many "readers" in a class, but the output of the students' rating scores would need to be calibrated for rater effects. To present scores adjusted for rater severity, the instructor can run a Rasch analysis on the collected data. This also formulates a summarized class-wide result for comparison. Accompanied by the feedback data, this information can then be shared with each student writer to create further opportunities for learning as an audience of readers. Sharing of class-wide scoring results can be done anonymously by using a coding system.

Rasch analysis

Rasch analysis is a psychometric tool widely used by researchers in SLA to measure constructs such as performance and motivation. It is also used in reliability assurance of language proficiency testing. The Rasch model (Rasch, 1961) and many-facet Rasch measurement (Linacre, 1989) allow for the facets of a construct to be simultaneously calculated and calibrated along a common interval scale, probabilistically accounting for the variance in human judgments. For example, when a student rates a peer's essay on a set of criteria, the thresholds between points on the rating scale, in each criteria, may change in meaning between criteria and between members of the rater group (Eckes, 2015). A score of eight on one criterion by a rater may be of a different severity than an eight by the same rater on another criteria. The Rasch model predicts this variability to a degree and estimates a *fair score* in comparison to an averaged rater severity (see Linacre, 2020b, p. 130). In the present design, the facets are writing performance, rater severity, and criteria.

An advantage of Rasch analysis is that values for 'rater severity' and 'good writing' can be calculated even if not all raters assess all essays, so long as the structure, or judging plan, underlining the peer exchange ensures sufficient interconnection between raters and essays. The design of the judging plan and the consistency of ratings highly affect the accuracy of overall measurement. Research has shown that resources applied to rater training can improve aspects of rating consistency, but that the consistency between raters is difficult to fully attain (Farrokhi et al., 2012; Weigle, 1998). It is an important aim of a Rasch analysis to achieve an accuracy of scoring appropriate to its purposes. The related fit statistics and the judging plan design are addressed further in the next section. Applying Rasch analysis within a peer-assisted learning process can place L2 writers in a position to

- assess and give feedback on a wider selection of essays written by peers,
- interpret fair scores from Rasch in conjunction with qualitative feedback comments,

- self-assess their own rater consistency and revisit their assessments of essays, and
- evaluate and reflect on the experiences of writing, reading, and reviewing.

The resulting cognitive conflict experienced and the assimilation of possibly conflicting information is thought to encourage learning.

The present study

This exploratory study critically evaluated an application of this approach in a semester-long academic course with high-proficiency L2 learners of English. The aim of the approach is to benefit student learning. To assess its success, the study seeks to answer the following questions:

1. Will this peer feedback and assessment design reliably meet its theoretical aims?
2. How will the participants perceive and engage in this process featuring anonymity and multiple roles taken at different stages?
3. Based on these outcomes, how might the design be improved upon for future applications?

Method

Participants

A class of 15 L2 speakers of English, all in their second year of university, participated in the 10-week process. The participants' proficiency level was CEFR B2 and above based on the general requirement of having a TOEIC score of 750 points to enter the course. Their proficiency was adequate for the set course book: *HBR's 10 Must Reads on Leadership* (Harvard Business Review, 2011), a book containing a selection of pragmatic and readable articles from thought leaders in the world of business. Two thirds of the participants were international participants coming from a variety of countries in Asia. Of the Japanese participants, all had educational experiences of at least 2 years outside of Japan.

Materials

Judging plan of an incomplete block design

Each participant rated and gave written feedback to essays assigned according to a judging plan. The judging plan of an incomplete block design allows for a peer-exchange procedure to be economically executed without each participant needing to rate every paper. 'Incomplete' means that each participant's essay is rated by varied subsets of classmates within the group (Eckes, 2015). As long as the raters and samples within the block sufficiently overlap, Rasch analysis will be able to produce group-wide measures (Linacre & Wright, 2002). Without this technique, group measures would require each rater to rate all essays. Such rating would be laborious and could raise concerns over feasibility and rater fatigue. Table 1 demonstrates this particular incomplete block design, which uses a repeating pattern to assign connections between participants and papers within the data set. Additionally, the exemplar paper was assigned to all raters and the instructor was assigned to rate each participant's work. Feedback comments and ratings were collected accordingly.

Table 1
Demonstration of Judging Plan

Student paper code / Student rater code		s-1	s-2	s-3	s-4	s-5	s-6	s-7	s-8	s-9	s-10	s-11	s-12	s-13	s-14	s-15	t-1
		r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code	r-code
s-1	p-code						7				6			8		9	7
s-2	p-code							8	7						8	7	7
s-3	p-code	8	9					7		7							7
s-4	p-code	8		8				9			9						8
s-5	p-code	7			8				7	8							8
s-6	p-code	7				7			8		9						7
s-7	p-code		9	5						6	8						7
s-8	p-code		7		9							7	9				8
s-9	p-code		7			6						4		10			8
s-10	p-code			7	10							7			9		9
s-11	p-code			8		7							7			8	7
s-12	p-code				8	4						3		9			8
s-13	p-code						3	5					7		7		6
s-14	p-code						1		5				7			8	6
s-15	p-code						3			5				8	7		6
ex-1	p-code	8	9	7	8	10	10	8	8	8	7	5	7	9	9	7	8

Note. *Incomplete block pattern (6 x 10, boxed with thick lines) repeated to create the plan. Same incomplete block design used for each criterion.*

Criteria for assessment

The participants rated on three criteria: Value to Reader, Clarity of Concept, and Language Mechanics, each on a 10-point Likert scale. The first criterion of Value to Reader is reflected in Topping's (1998) definition of peer-assessment. The categories, or levels, of each criterion were not given detailed descriptors. The simple design of three criteria was to help the participants remember the criteria and transfer learning onto their own writing. Measuring value to the reader introduced a degree of subjectivity because readers can hold different senses of value. Nonetheless, the set of criteria as a whole was meant to be straightforward to complete and to help facilitate accompanying feedback comments. The criteria and scale were discussed in an orientation session where participants were encouraged to use the full range of the scale as they deemed appropriate. The participants were not given specialized assessment training. Unpredictable ratings by participants were anticipated as becoming material which peers could later critically evaluate. Both narrow and unpredictably wide use of the rating scales beyond a certain range was understood to possibly lessen the precision of the Rasch analysis (Eckes, 2015).

Procedure

The workflow had participants write, evaluate, and make interpretations. The instructor gathered data, conducted the Rasch analysis, and prepared a summarized report on the analysis and feedback comments for the participants. The scheme took place over 16 sessions, spanning 10 weeks. In brief, each participant

1. covered course materials;
2. completed the writing task;
3. assessed and gave feedback on a selection of writings from peers (quantitatively and qualitatively, assigned via the judging plan);
4. received an analysis and feedback report prepared by the instructor, with the instructor's feedback added anonymously to the report;
5. made interpretations from the experience and from the results of the report; and
6. completed a reflection paper.

The writing task was a single 1000-word argumentative essay centering on a topic raised by articles in the course book focusing on leadership in business. Participants had to critically address conflicting aspects of the articles and introduce counterarguments using sources as appropriate. They were to make a claim and support it, that is, not to merely summarize the content of articles. The participants kept their essays private in the early stages to maintain anonymity.

One exemplar paper, coded p-E, was included anonymously from a previous course. It was chosen for having above average language mechanics but with a strongly worded, controversial argument. Each participant was assigned this exemplar paper in addition to a subset of four anonymized essays written by peers according to the judging plan demonstrated in Table 1. Therefore, the participants each had a total of five essays to evaluate along three criteria. Each participant received PDF files of their assigned essays and completed an online form from home. No minimum wordcount or other requirements were requested of the participants. The evaluations were collected, anonymized and kept in a password-protected file offline. Teacher assessment and written feedback for all papers was included anonymously into the data at this point.

The analysis and feedback report were prepared by the instructor for the participants. Calculations and the output of data were done with Facets software, version 3.83.2 (Linacre, 2020a). The data produced by this software is typically meant for research purposes, so the instructor simplified this output for the participants. The report consisted of the following:

- vertical rulers (seen in the results section)
- simplified fit statistics
- tables of unexpected responses
- simple explanation and instructions
- the feedback the participant's essay received
- all feedback received by the exemplar essay

The Rasch analysis included four sets of results: one for all criteria combined and one each for the three criteria. Participants were given a unique ID code to locate their results in relation to the anonymized essay and rater data. The identity of each participant was protected within the report. Participants also received anonymized feedback comments for their paper. Participants did not revise their original papers; rather, each participant completed a reflection paper on their experiences at the different stages during the process.

Reflection

The participants wrote a 500-word reflection paper to bring together and reinforce their learning, as underpinned by Bloom's taxonomy (Anderson & Krathwohl, 2000; Bloom, 1956). Writing the reflection paper invited the participants to once again learn by assessing (Topping, 1998). This self-assessment gave the participants an opportunity to review the ratings and feedback they had given and received, evaluate

conflicting experiences, and reflect. The reflection paper also served as data for the researcher to determine the efficacy of the procedure.

A final step was to anonymously complete an exit survey with the following five statements, each rated on a 6-point agreement scale:

1. Giving feedback and ratings to papers written by my peers required deep thinking.
2. Giving feedback and ratings helped me better self-assess my own writing.
3. It was good that the peer-exchange was anonymized (names were kept secret).
4. I can better see how responses by readers can be unpredictable or unexpected.
5. The final analysis and feedback report were helpful.

Analysis

The analysis focused on aspects of the feedback comments, the class-wide measures, and the participants' perceptions about the procedure. The volume and variety of feedback comments were analyzed in relation to the effects of using the judging plan. The model statement was set before running Facets software, version 3.83.2 (Linacre, 2020a; also see Linacre, 2020b, for further examples and explanation of model statements). The raters and the essays were set to the rating scale model. The facet of criteria was set to the partial credit model, because it was assumed that use of rating scales would vary by criteria (Value to Reader, Clarity of Concept, and Language Mechanics). Attention was paid to underfitting (unpredictable) ratings where raters erroneously rate less-effective papers leniently or rate highly rated papers harshly. Attention was also paid to overfitting raters, those who either used a narrow range on the Likert scale or those who potentially exhibited a halo effect and scored the criteria holistically. General impressions of the participants' perceptions about the procedure were analyzed by examining the results of the anonymous exit survey and by highlighting selected excerpts from the reflection papers.

Results

Descriptive statistics on feedback

A summary of the amount and distribution of feedback comments indicated that the judging plan ensured that each paper received a reasonable amount of feedback despite the wide variation in feedback length given by participants. To recap, each participant wrote a 1000-word argumentative essay and a 500-word reflection paper. No word requirements were imposed on feedback. Even so, most participants voluntarily wrote a sizable amount in relation to the writing assignments.

Each participant wrote on average 292.6 words (*SD* = 202.2 words) of feedback in total for their five assigned papers. The high variation of feedback given by the participants was balanced out via the structure of the judging plan. As a result, the standard deviation for the feedback each paper received was halved (*SD* = 102.7 words). Each paper received a mean amount of 241.2 words in total feedback from the four classmates assigned as reviewers. This was in addition to teacher comments, which were written anonymously and roughly matched in length. Reasons for variation in participant feedback are unclear, but the data indicates that the judging plan effectively served to balance out the effects of participant variation and provide a base level of feedback for each participant's essay.

A portion of the variance in feedback for each essay appears to be accounted for by a moderate negative relationship between the evaluation of essays and the word counts of feedback (r = -.529, p = .047), which would indicate that essays with lower evaluations tended to receive more written feedback than more highly rated essays. The evaluation of each paper was estimated using its adjusted fair score via Rasch measurement. The exemplar essay, reviewed by each participant, received a total of 771 words of feedback.

However, there was a range of feedback for this essay, with an average of 55 words produced by each participant (n = 14 participants giving feedback, *SD* = 53.5 words) and one participant giving no feedback. Put all together, however, this collection of feedback provided all participants a common reference to review and re-evaluate within the analysis and feedback summary report.

Rasch measures and fit statistics

The "vertical ruler" in Figure 1 is a graphic display that shows logit measures for the three criteria combined and it is indicative of the criterion-specific vertical rulers also in the participants' feedback and assessment report. For value, concept, language, and all criteria combined, the rating data is transformed along a common equal-interval measure of logits, or "log-odds units" (see Bond & Fox, 2015, p. 46). The left side of Figure 1 shows participant ability (writing performance) ranked from high to low (top to bottom), with argumentative essays (papers) P, C, and N receiving the highest adjusted scores. The next column represents rater severity (with leniency at the bottom), where raters #12, #11, and #2 appeared to be the strictest raters. Comparing laterally, the vertical rulers line up participant ability with rater severity. While each rater rated only five papers, the class-wide approximation in Figure 1 tells us that Rater #14, for example, would have probably rated half of all essays more positively and the other half more negatively. Papers J and B would have a 50/50 chance of being rated high or low by Rater #14. Such comparisons imply a rough match at this point between performance and severity.

However, the logit spread observed for rater severity (3.39 logits) exceeds the logit spread for writing performance (2.82 logits). The severity and lenience of Raters #7 and #12 may have been exaggerated due to too many raters using a narrow range of the scales for each criterion. The category threshold measure on the right side of the vertical rule demonstrates that many papers received scores of 7 and 8 across, particularly for the criteria of Value. This narrow use of the scale reduces the accuracy of measurement. Use of the whole scale was discussed with participants in class, but in the end, the rating patterns seemed to mimic the typical school grading pattern A-D and F. This instinctive use of a 10-point scale is natural and should have been better foreseen in the design. Research suggests that a scale of six categories or less can ensure better measurement by minding potential limits in raters' motivation and working memory (see Nemoto & Beglar, 2014).

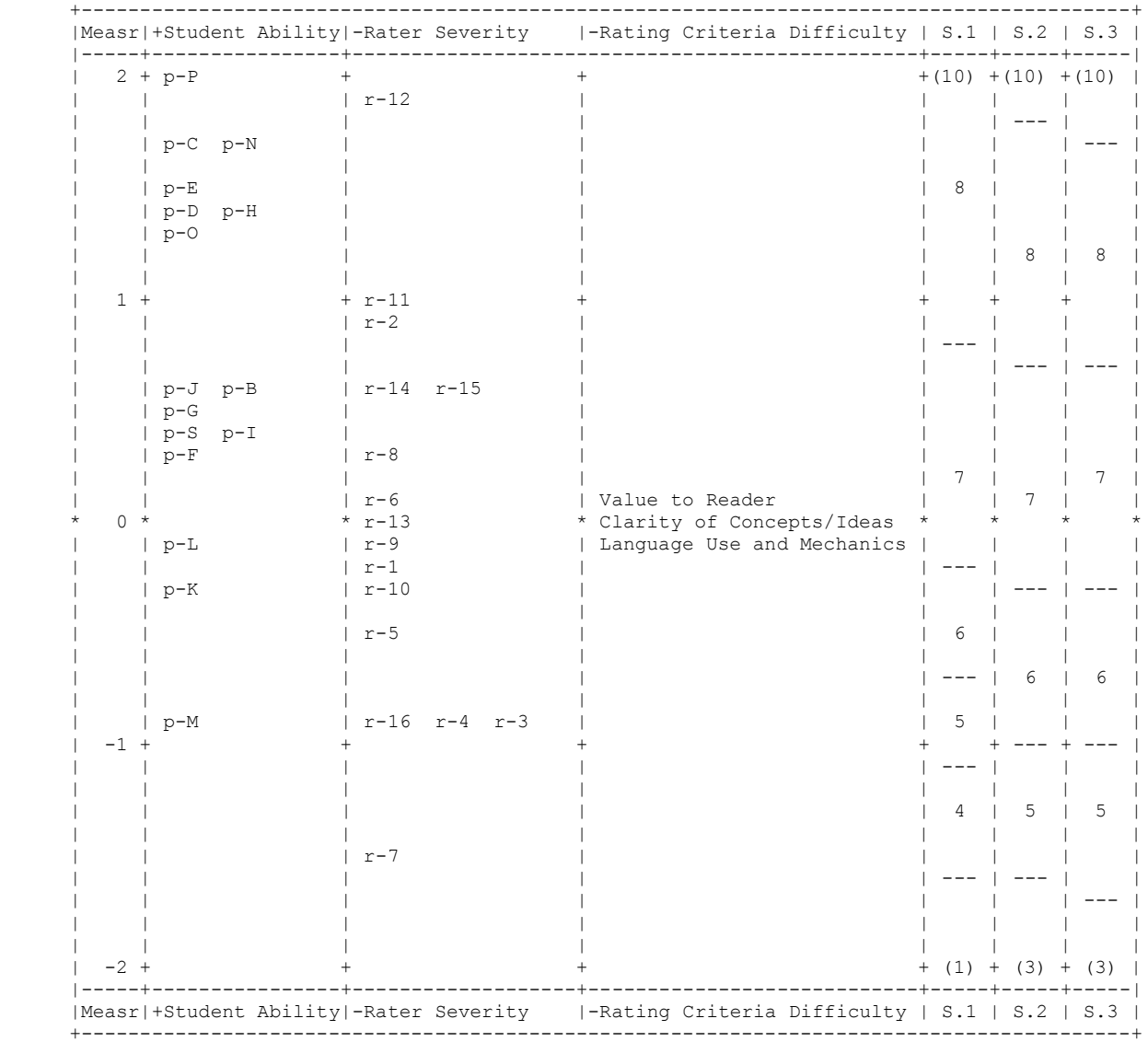

Figure 1. Vertical Rulers for All Three Criteria. Right-side column labels: S.1 = Value to Reader, S.2 = Clarity of Concept, S.3 = Language Mechanics.

Regarding fit of the data, one third of the participants fit the Rasch model (Table 2, bolded), while almost half of the participants overfit, likely giving many essays common scores of 7 and 8. A few participants underfit by giving unpredicted responses such as assigning highly scored papers low scores, or vice versa. Table 2 also shows that the infit and outfit measures by each participant varied by criterion (Table 2, fitting between MnSq 1.00+/-.50 in bold by criterion). The fit range of MnSq 1.00+/-.50 for each rater and paper is most productive to the model's overall measurement (Linacre, 2020b, p. 286). Outlying ratings were predicted to become a resource for participants to evaluate. However, data overfit was more than anticipated and this reduced the accuracy of measure. Scores of 1-5 points were rarely given. In an attempt to narrow the logit range of rater severity, the data was re-analyzed on a 7-point scale, with categories 1-4 combined. A paired sample *t*-test saw marginal improvement, but not at the level of significance. No outlying data was removed because this was not a summative assessment and because

each participant's data was part of the analysis and feedback report. For more precise measures, the papers would need to be reassessed using a 6-point scale, for example, as outlined by Nemoto and Beglar (2014). For the current results, the separation of assessment for each criterion was relatively low for both raters and papers (2.29, 2.69, 1.51 and 2.19, 2.35, 1.79, respectively) due to the narrow use of the scales. This indicated that raters and papers were not sufficiently separated into groups of 3 or more. Nonetheless, the data was able to show a low-resolution separation of performance and severity. The participants were informed that finer positions between papers versus raters are not as accurate.

Table 2
Fit Statistics for Rater Severity

	Infit	Infit by Criteria			Outfit	Outfit by Criteria		
Rater	Combined	Value	Concept	Language	Combined	Value	Concept	Language
r-1	3.88	2.21	3.08	3.84	3.91	2.81	2.93	3.49
r-2	2.04	1.91	1.99	2.78	1.98	1.93	1.95	2.11
r-3	1.74	1.65	2.98	1.53	2.25	2.88	3.90	1.86
r-4	**1.46**	1.92	0.42	2.00	**1.47**	1.93	0.52	2.01
r-5	**1.22**	0.47	**1.13**	1.00	**1.21**	0.49	**1.10**	**1.00**
r-6	**1.07**	**0.81**	**0.94**	1.88	**1.19**	**1.00**	**0.97**	2.10
r-7	**0.90**	**0.71**	**1.08**	**0.53**	**0.85**	**0.74**	**1.01**	**0.52**
r-8	**0.57**	0.33	**0.50**	**0.90**	**0.57**	0.42	**0.51**	**0.88**
r-9	**0.54**	0.29	**0.71**	0.43	**0.57**	0.37	**0.72**	0.43
r-10	0.48	**0.71**	0.27	0.42	0.45	**0.67**	0.27	0.38
r-11	0.47	**1.45**	0.38	0.14	0.48	**1.45**	0.37	0.13
r-12	0.45	0.28	0.19	**1.08**	0.45	0.23	0.17	**1.02**
r-13	0.45	0.08	**0.77**	0.20	0.44	0.12	**0.77**	0.18
r-14	0.35	0.34	0.19	**0.61**	0.32	0.34	0.19	**0.52**
r-15	0.33	0.35	0.20	0.19	0.31	0.47	0.20	0.22
r-16	0.14	0.05	0.14	0.15	0.13	0.06	0.15	0.12
Mean	1.01	0.85	0.93	1.11	1.04	0.99	0.98	1.06
SD	0.94	0.73	0.95	1.06	0.98	0.93	1.07	0.98

Note. Ordered by the infit figures of the three criteria combined (1.00+/-.50 MnSq bolded). Participant group of 15 members, one instructor. Criteria = partial credit model.

The fit statistics for the papers revealed a marked result for the exemplar paper, coded p-E. As predicted, the fit statistics for p-E indicated unpredictable responses for Value to Reader (Infit MnSq = 1.99, Outfit MnSq = 2.17) and were the only underfitting figures for value among the papers (Table 3). More variation in responses could be because it was rated by all participants. It could also be due to it garnering a wider range of reactions to the paper's somewhat strongly stated arguments. The participants appeared to assess the value of p-E in different ways and this was also borne out in the essay's qualitative feedback comments as well. Both the comments and rating results for exemplar p-E were provided to all participants as an informative subsection of the analysis and feedback report.

Table 3
Fit Statistics for Participant Ability (Performance of Papers)

	Infit	Infit by Criteria			Outfit	Outfit by Criteria		
Rater	Combined	Value	Concept	Language	Combined	Value	Concept	Language
p-A	1.86	**0.58**	2.35	2.93	1.96	**0.83**	2.39	3.08
p-B	1.81	**0.83**	**1.33**	2.51	1.75	**0.70**	**1.28**	1.98
p-C	**1.55**	**0.94**	**1.10**	**1.11**	1.84	**0.72**	**1.14**	**0.99**
p-D	**1.28**	**0.97**	0.42	**0.85**	**1.45**	**1.09**	**0.55**	**0.84**
p-E	**1.25**	1.99	1.86	**1.07**	**1.27**	2.17	1.94	**1.03**
p-F	**1.06**	**1.14**	**1.24**	**0.86**	**1.10**	**1.35**	**1.21**	**0.95**
p-G	**0.81**	**0.86**	**0.79**	**1.42**	**0.85**	**1.05**	**0.82**	1.57
p-H	**0.80**	**0.82**	**0.61**	**0.80**	**0.80**	**0.81**	**0.62**	**0.80**
p-I	**0.75**	0.40	**0.77**	1.89	**0.77**	0.41	**0.78**	1.96
p-J	**0.69**	**0.67**	0.16	0.30	**0.63**	**0.78**	0.17	0.27
p-K	**0.68**	**0.71**	0.14	**0.54**	**0.67**	**0.70**	0.15	**0.54**
p-L	**0.53**	0.20	**0.69**	**0.80**	**0.51**	0.19	**0.70**	**0.73**
p-M	**0.50**	0.11	0.46	**0.54**	0.49	0.09	0.47	**0.54**
p-N	0.37	0.43	0.16	0.19	0.38	0.46	0.18	0.19
p-O	0.35	0.33	0.46	0.04	0.35	0.34	0.44	0.04
p-P	0.21	0.26	0.20	0.16	0.21	0.26	0.21	0.15
Mean	0.91	0.70	0.80	1.00	0.94	0.75	0.82	0.98
SD	0.51	0.46	0.64	0.83	0.56	0.51	0.65	0.81

Note. Ordered by the infit figures of the three criteria combined (1.00+/-.50 MnSq bolded). Fifteen papers written by the participants, one exemplar. Criteria = partial credit model.

The anonymous exit survey attempted to measure how the participants perceived different aspects of the peer-exchange process (Table 4). The results for the 6-point scale agreement items indicated that most participants could strongly endorse the five statements. However, the minimum and maximum values show that participant sentiment was not uniform. Scores of 2 and 3 showed that a few participants disagreed with at least some of the intentions of the design.

Table 4
Exit Survey Results

	Q1	Q2	Q3	Q4	Q5
Replies	15	15	15	15	15
Mean	5.00	5.07	5.13	5.07	5.13
Std. Deviation	1.00	0.88	0.92	0.88	0.92
Skewness	-1.98	-0.86	-0.94	-0.14	-0.94
Kurtosis	5.68	0.67	0.52	-1.78	0.52
Minimum	2.00	3.00	3.00	4.00	3.00
Maximum	6.00	6.00	6.00	6.00	6.00

Note. 6-point scale (1 = strongly disagree, 6 = strongly agree).

Discussion

Theoretical aims

The act of assessing and giving feedback, as in Topping's (1998) view, seemed effective for learning in the Piagetian sense where each participant could encounter *cognitive conflict* between what they think they know and new experiences. The learning environment was student-centered and it provided the participants opportunities to better resolve conflicting information independently. The data generated suggested that the participants were learning by assessing at three stages of the exchange procedure. The first stage was when they assessed and gave feedback on their peers' papers. The second was when they had to evaluate the measures and qualitative feedback of various types provided in the report. Finally, the reflection paper provided a third opportunity for the participants to consolidate their learning and re-assess their performance.

The participants successfully completed each task and submitting work on time at each stage and this helps confirm that the process was straightforward for the students to follow. To attempt this peer exchange approach in another context, the instructor would need the organizational skills to manage the exchange of data and the knowledge needed to use Rasch analysis software (see e.g., Bond & Fox, 2015; Eckes, 2015; Linacre, 2020b). Attempting to transfer this peer feedback and peer assessment design to other learning situations would require adjusting the procedure accordingly. The length and number of tasks to complete could be carefully adjusted based on learner proficiency and learning aims. Writing tasks other than essay writing are possible. One example could be to have students produce and exchange a set of pragmatically written emails. In any case, the design generally appears to be applicable to writing development in a language course given proper preparation. This includes the use of a judging plan to enable a manageable and balanced exchange of varied feedback and assessment ratings.

The participants took on the role of the distant reader, as viewed by Booth et al. (2008), but the results of the fit statistics showed that the consistency of their assessments was less than expected. For a standardized testing situation, such evaluation would not be acceptable, but there was much evidence that a mechanism for learning was achieved within this peer exchange environment. It was key for the participants to interact with their peer's writing, the rubric, and the subsequent feedback information during the procedure. The participants saw the varying interpretations of value, especially in the case of paper p-E. Generally, the ratings by the participants diverged at a finer scale, but the exit survey showed high engagement, growth in critical thinking skills, and improved self-awareness in the writing process. Pedagogical value could then be seen in formative aspects of writing development rather than summative testing.

Perceptions by participants

Through observation of the participants' work and in-class discussion, it was clear to the researcher that the approach was positively received. This impression about the participants' perceptions and engagement was supported by the group's effortful qualitative comments in many of the reflection papers. In particular, the participants seemed to value anonymity as a way to better embody the role of a reader, apart from social distraction. One participant could experience how writers sometimes structure arguments in ways that are difficult for readers to receive.

> I was able to understand how other people write their essays (good and bad) and from those problems such as unclear structure, tone used in essay as well as grammar misusing and so on, I managed to understand how it feels when seeing these problems as a reader. I was like a lost cat in the forest could not find the way out. I assume this is what happens

> to those people who were evaluating my essay, too. It must be very difficult and hard. – Participant A

Another participant reflected on the conflict within the evaluation process.

> I honestly feel happy to get those mixed feedbacks from my classmates and the scoring because it encouraged me to improve my learning ability on logical and creative thinking for understanding the articles. Exchanging feedbacks in this course has not only helped me to develop self-awareness of my writing ability, but also helped me to know how to provide value to the audience with our creative thinking and writing skill. – Participant B

While the participants generally seemed to find the process insightful, a few participants shared a preference for teacher feedback. Interestingly, the teacher's comments were in fact provided anonymously within the analysis and feedback report. Each participant also received additional and open feedback after the process from the instructor. This raises interesting questions about the perceptions of feedback, its types, and its sources. This could be an area for further investigation.

Future considerations

To remedy the issues about accuracy of measurement, the design of the rating scale would be the first consideration. The 10-point scale in this context too closely resembled a typical school grading system, leaving half of the Likert scale (points 1-5) not being utilized by most participants. The rating patterns of most participants saw many 7s and 8s (analogous to C and B grades) and almost no scores below 6 points for the three criteria. This narrow use of the ratings scales likely exaggerated the unpredictability of some other raters. Such measurement error could be avoided by choosing a point scale that more intuitively encourages broader scoring patterns, which leads to more productive ratings for Rasch analysis. This would be a highly economical solution instead of other possible strategies like time spent on rater training. Even with time and robust resources, rater training can yield limited benefits (Eckes, 2015; Farrokhi et al., 2012; Weigle, 1998). Nemoto and Beglar (2014) outlined how a 6-point scale would avoid over-stressing the working memory of raters by applying descriptors for a smaller number of categories on each scale. An even-numbered scale would also prevent participants from making neutral assessments. By simplifying the scale, learners may be able to better interact with peer work clearly through the rubric, even if interpretations vary. This interaction can become an improved mechanism for independent learning in addition to the receiving of feedback and assessment. If applied in other L2 learning contexts, the criteria of a rubric with simplified scales could be adjusted according to the task. For the procedure as a whole, considerations for L2 proficiency, writing or speaking task, and learning goals would need to be made carefully.

Conclusion

Anonymity and Rasch analysis via a judging plan were successfully employed in a peer feedback and assessment design to facilitate the aims of building awareness of reader needs and the ability to resolve differences among a variety of feedback and assessment. These features provided all participants in the group with class-wide measures on writing performance and peer rating patterns. The results of the study showed that the judging plan also helped ensure that the variation of feedback given by individuals was balanced out and allowed for a variety of peer-feedback for each member of the group. Participant reflections and exit survey responses indicated positive reception by the participants and that the goal of creating learning opportunities for participants at multiple stages was met. However, the review of the Rasch analysis figures raised concerns over lower-than-expected accuracy of the measure due to a narrow use of the rating scales. A 6-point rating scale for each rating criterion is proposed in order to better elicit

discerning scoring responses by peer raters. The results of this single application cannot be easily generalized beyond the context of these learners. However, its evidence of peer-assisted learning and the feasibility of its improvements suggested that this design for peer feedback and peer assessment with L2 participant writers is worth further investigation.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Pearson.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook 1: Cognitive domain*. Longman.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Booth, W. C., Colomb, G. G., & Williams, J. M. (2008). *The craft of research* (3rd ed). University of Chicago Press.

Cheng, K.-H., & Tsai, C.-C. (2012). Students' interpersonal perspectives on, conceptions of and approaches to learning in online peer assessment. *Australasian Journal of Educational Technology*, *28*(4), 599–618. https://doi.org/10.14742/ajet.830

Eckes, T. (2015). *Introduction to many-facet Rasch measurement*. Peter Lang.

Falchikov, N. (2001). *Learning together: Peer tutoring in higher education*. Routledge.

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. https://doi.org/10.37546/JALTJJ34.1-3

Foot, H., & Howe, C. (1998). The Psychoeducational Basis of Peer-Assisted Learning. In K. J. Topping & S. W. Ehly (Eds.), *Peer-assisted learning* (pp. 29–39). Routledge.

Garner, B. A. (2012). *HBR guide to better business writing*. Harvard Business Review Press.

Harvard Business Review. (2011). *HBR's 10 must reads on leadership*. Harvard Business Review Press.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2020a). *Facets*® (3.83.2) [Computer software]. https://winsteps.com

Linacre, J. M. (2020b). *A User's Guide to Facets Rasch-Model Computer Program*. winsteps.com.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*(4), 486–512.

Liu, J., & Hansen, J. G. (2002). *Peer response in second language writing classrooms*. University of Michigan Press.

Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279–290. https://doi.org/10.1080/13562510600680582

Lundstrom, K., & Baker, W. (2009). To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing*, *18*(1), 30–43. https://doi.org/10.1016/j.jslw.2008.06.002

Nemoto, T., & Beglar, D. (2014). Developing Likert-Scale Questionnaires. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings* (pp. 1–8). JALT. Available online: http://jalt-publications.org/files/pdf-article/jalt2013_001.pdf (accessed on 18 November 2020).

Piaget, J. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*. University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Volume IV: Contributions to biology and problems of medicine* (pp. 321–333). University of California Press.

Rotsaert, T., Panadero, E., & Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: Its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education*, *33*(1), 75–99. https://doi.org/10.1007/s10212-017-0339-8

Topping, K., & Ehly, S. (1998). *Peer-assisted Learning*. Routledge.

Topping, K. J. (1998). Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research*, *68*(3), 249–276. https://doi.org/10.3102/00346543068003249

Vygotsky, L. S. (1978). *Mind in society*. Harvard University Press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.

Xu, Y., & Liu, J. (2010). An investigation into anonymous peer feedback. *Foreign Language Teaching and Practice*, *3*, 44–49.

Yu, S., & Lee, I. (2016). Peer feedback in second language writing (2005–2014). *Language Teaching*, *49*(4), 461–493. https://doi.org/10.1017/S0261444816000161

Zamel, V. (1982). Writing: The process of discovering meaning. *TESOL Quarterly*, *16*(2), 195. https://doi.org/10.2307/3586792