Many-faceted Rasch Analysis of Student Peer Assessment

Trevor A. HOLSTER

福岡女子大学文学部・国際文理学部紀要「文藝と思想」
第76号　pp.69～86 抜刷　2012年2月

福岡女子大学文学部・国際文理学部紀要
「文藝と思想」第76号　2012年2月　69～86頁

Many-faceted Rasch Analysis of Student Peer Assessment

Trevor A. HOLSTER

Abstract

Peer assessment of language performance is a potential catalyst for learning by teaching, where peer raters improve their own proficiency by rating and mentoring other students, but issues of rater performance must be addressed to ensure that these benefits are achieved. Many-faceted Rasch measurement can provide quality control of rater performance, as well as more traditional analysis of item performance and interval level measures of proficiency, but the very large quantities of data generated by peer assessment are a major practical limitation. A peer assessment module was developed for the open-source MOARS audience response system and used to gather data from 185 sets of ratings by both peers and teachers. The system addressed the practicality of data collection and peer ratings overall showed adequate agreement with teacher ratings for low stakes classroom use, but the performance of individual raters was found to vary greatly, supporting the importance of Rasch measurement to adjust for rater performance and provide diagnostic analysis of rater performance.

Background

Standardized tests such as TOEFL (ETS, 2008) aim to measure proficiency by isolating characteristics of tasks sampled from the relevant domain and then

producing test items that incorporate the key characteristics of the real world tasks (Chapelle, 2008). Such tests have great utility as they focus on stable traits that can be tested cheaply using multiple choice tests. They can also provide separate diagnostic scores for different traits, allowing students and teachers to identify strengths and weaknesses and tailor study accordingly. However, many testing experts also support performance tests, which integrate multiple skills into a single overall performance. For example, writing an essay requires integration of things such as vocabulary, grammar, rhetorical structure, and content knowledge into a single piece of work. A well documented difficulty with performance assessments is the need for human raters, requiring monitoring of rater performance. McNamara (1996) and Weigle (1994) provided seminal accounts of the use of many-faceted Rasch measurement (MFRM) (Linacre, 1994) to check that raters are behaving consistently and adjust for differences in rater severity.

While performance tests can be used to give summative scores, such as assigning a course grade, they are also argued to have formative benefits, aiming to guide study behavior or teaching (Hughes, 2003). One major benefit of Rasch models over classical test theory (CTT) is the provision of fit statistics that show whether students' performances are consistent with the overall group (Bond & Fox, 2007). Analysis of misfitting students can help teachers identify students who need remedial study, and Engelhard (2009) extended this to show how MFRM can be used to check that students with disabilities are not disadvantaged in tests.

As well as using teacher ratings from performance assessment to give formative feedback to students, peer assessment and self assessment are potentially valuable resources for both learning and assessment, and peer assessment using MFRM has been used in Japanese high schools and universities (Fukuzawa, 2010; Negishi, 2010; Saito, 2008). Topping (2005) discusses "learning by teaching", where students evaluate other students' performances and then give feedback and provide coaching, forcing the peer coaches to think more deeply about what a good performance is and how to communicate this to others. However, large scale peer assessment, regardless of whether it is intended for summative or formative purposes, requires very clear objectives so that all teachers and students understand what is expected. This is then operationalized as a scoring rubric and Rasch fit

statistics allow detailed diagnosis of problems in interpreting this. Misfitting raters may be misunderstanding the rubric or employing it idiosyncratically, casting doubt on misfitting peer raters' ability to benefit from learning by teaching. Thus, MFRM allows measurement and diagnosis of students' performances of the task itself, and also of their understanding of the performances of others.

Instrument and Data Collection

A difficulty with peer assessment is data collection, so a peer assessment module was written for the open source MOARS audience response system (Pellowe, 2002, 2010). This allows students to rate peers using an internet browser with an interface simple enough to be displayed as a web page on a mobile phone. Teachers can export formatted data ready for MFRM analysis, making it practical for classroom teachers to quickly analyze peer assessment results (Holster & Pellowe, 2011). A school wide trial of the MOARS peer assessment module was carried out in the first semester of 2011 in presentation skills classes, so a simple rubric was developed for use by both teachers and peer raters. Of the sixteen classes in the program, comprising 232 students, six were taught by Japanese teachers of English (JTEs) and ten by non-Japanese teachers. Two of the three Japanese teachers declined to try the peer assessment system, leaving a sample of 172 students in 12 classes taught by one Japanese teacher and five non-Japanese teachers.

The grading rubric used for the first presentation is shown in Appendix 1, covering 12 items graded on a 4 point scale from "A" to "D". This was administered using the MOARS system to collect data for a 3 faceted Rasch analysis. Classroom teachers are implicitly familiar with a 2 faceted measurement model, where only items and persons are considered. Rasch analysis of such a dataset models the probability of a successful response to an item as a function of the difference between the facet of person ability and the facet of item difficulty (Bond & Fox, 2007). However, data from judged performances introduces raters who may differ in severity, requiring the facet of rater severity to be included (Linacre, 1994; McNamara, 1996). Thus, the probability of success can be 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. Thus, the odds of success increase with greater person ability, but decrease with greater item difficulty or greater rater severity.

Following the pilot round of presentations, the rubric was simplified and reduced to nine items for the second presentation, as shown in Table 1. Teachers felt that "Confidence" was not well defined and was redundant, while the three content items were extremely easy and contributed little to measurement. Therefore, "Confidence" was removed entirely and the three content items were collapsed into a single item called "Content".

Because peer raters only rated performances by students within their own class, it was necessary to link the resulting disjoint subsets of data. This was achieved in the first presentation by making three training videos of teachers making practice presentations, one being deliberately very poor, one mediocre, and one good in order to gather data across the full range of the rubric. Students rated these in class to familiarize them with the rubric and rating sheets and to gather responses for linking. These linking videos allowed the severity of all raters to be directly compared, allowing all class groups to be measured on a common scale. Linking in presentation two was further improved by asking teachers to rate video recordings of student presentations, thus allowing teacher raters

Table 1 Pilot versus Revised Rubric Items

	Pilot Items	Revised Items
Non-verbal communication	Confidence	–
	Notes/Reading	Using notes
	Eyes	Eye contact
	Hands	Gestures
	Body	Posture/movement
Voice	Speed	Speed
	Volume	Volume
	Pausing	Pausing
	Intonation	Intonation
Content	Organization	Content
	Relevance	
	Completeness	

to be directly compared and peer raters to be linked through teachers.

Results

As the effectiveness of the rubric and teacher raters was a prerequisite for any further analysis, these were evaluated first. The severity of the teacher raters was measured using the complete dataset to provide anchoring values for subsequent analyses. Next, the performance of the teacher raters was analyzed in isolation. The results of this are shown graphically in Figure 1, with student ability, rater severity, and item difficulty mapped onto a common vertical logit scale on the left, with the raw rating scale on the right. The logit scale gives equal interval measurement, but it can be seen that the raw rating scale does not, with the interval between a rating of 2 and 3 much greater than between 1 and 2. Following standard Rasch measurement practice, mean item difficulty is set to an arbitrary value of 0.00 logits, but it is apparent that most students have a very high probability of success on all items, with "completeness", "organization", and "relevance" extremely easy for this sample of students. It is also apparent that there was a range of rater severity of more than 2.5 logits, an extremely large value. If a student had a 50% chance of success on an item with a rater of middling severity, they would have a 78% chance of success with the most lenient rater and a 22% chance of success with the strictest rater. Thus, raw scores from teacher raters are not interchangeable and adjustment for rater severity would be essential for high stakes decisions.

It can also be seen that Rater 1 (the author) rated some performances three times to compare intra-rater consistency, so is shown as Rater 1 for live ratings in class, and Rater 1b and Rater 1c for subsequent ratings. This rater taught and rated only relatively low proficiency groups, raising the question of whether raters unconsciously adjusted their ratings to use the full range of the scale, resulting in stricter ratings for higher proficiency groups and more lenient ratings for lower proficiency groups. If such adjustments are consistent, *Facets* can provide adjusted measures of student ability, but if raters perform inconsistently, then measurement will be degraded. Rasch fit statistics provide quality control of this, allowing

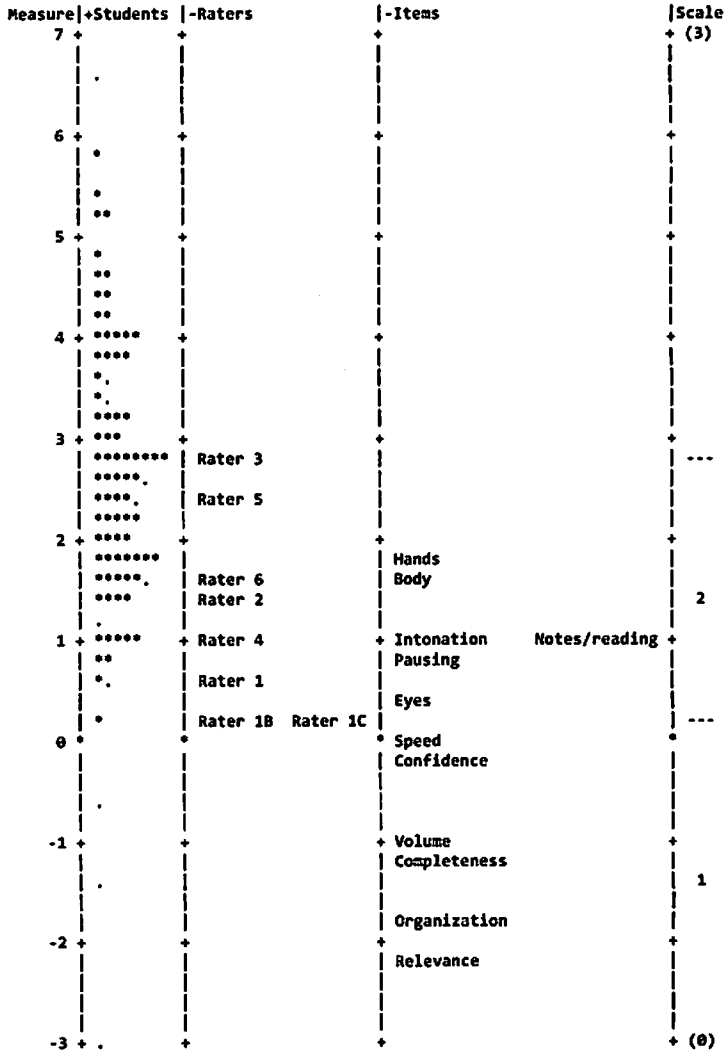

Figure 1. Facets measurement rulers for teacher raters. Student ability, rater severity, and item difficulty are mapped onto a shared measurement scale.

identification of inconsistent rubric items, raters, and students. Table 2 shows the measurement report for teacher raters, arranged by fit-to-the-model. Although Rater 1 was somewhat lenient, with a measure of 0.63 logits versus the mean of 1.41 logits, his infit mean-squared and outfit mean-squared statistics, 1.06 and 1.01 respectively were close to the expected value of 1.00. These figures indicate 6% more randomness for in-lying responses and 1% more randomness for outlying responses, levels far too small to threaten measurement. Rater 6 was the least consistent, with infit and outfit mean-squared statistics of 1.46 and 1.56 respectively, a level still contributing to measurement but becoming affected by noise. This rater is interpreting the rubric somewhat differently from the others, but analysis of this rater's responses in isolation found infit and outfit statistics of 1.01 and 0.98, respectively, indicating that the rater was self-consistent, and therefore able to evaluate students effectively for course grades.

The next step in the analysis concerned the performance of the items. A key assumption of Rasch models is that items measure a unidimensional trait, with misfitting items an important indicator of whether this precondition for measurement has been adequately met. Table 3 shows the measurement report for items, with mean values of both infit and outfit of 0.97, indicating slightly more

Table 2 Teacher Raters Measurement Report for Presentation 1

Total Score	Total Count	Obs Ave	Fair-M Ave	Measure	Model *SE*	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim Disc	PtM	Corr PtEx	Exact Obs %	Agree. Exp %	Rater
728	367	2.0	1.78	1.95	0.09	1.46	5.6	1.56	5.8	0.40	.57	.74	46.4	45.7	Rater 6
620	335	1.9	2.31A	0.63	0.10	1.06	0.7	1.01	0.0	1.03	.77	.73	50.2	41.1	Rater 1
757	374	2.0	1.63A	2.32	0.09	0.95	-0.6	0.95	-0.5	1.04	.76	.71	48.5	42.6	Rater 5
612	396	1.5	1.87	1.74	0.09	0.89	-1.6	0.90	-1.3	1.07	.63	.71	64.3	51.5	Rater 2
660	504	1.3	1.46A	2.78	0.08	0.84	-2.8	0.82	-3.0	1.16	.74	.71	36.8	38.0	Rater 3
284	143	2.0	2.45	0.23	0.14	0.84	-1.4	0.80	-1.8	1.25	.82	.66	46.8	41.2	Rater 1B
47	24	2.0	2.41	0.34	0.35	0.82	-0.6	0.76	-0.8	1.38	.87	.67	52.8	40.1	Rater 1C
739	383	1.9	2.05	1.30	0.09	0.72	-4.4	0.71	-4.3	1.35	.78	.67	50.5	44.5	Rater 4
555.9	315.8	1.8	1.99	1.41	0.13	0.95	-0.6	0.94	-0.8		.74				Mean (Count: 8)
238.4	144.8	0.2	0.35	0.88	0.08	0.22	2.8	0.25	2.8		.09				*SD* (Pop)
254.8	154.8	0.3	0.37	0.95	0.09	0.23	3.0	0.27	3.0		.10				*SD* (Sample)

Model, Pop: RMSE.15 Adj(True) *SD* .87 Separation 5.70 Strata 7.93 Reliability (not inter-rater) .97
Model, Samp: RMSE.15 Adj(True) *SD* .93 Separation 6.10 Strata 8.47 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 517.7 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.8 d.f.: 6 significance (probability): .34
Inter-Rater agreement opportunities: 983
Exact agreements: 465 = 47.3% Expected: 412.2 = 41.9%

predictable data than modeled. The least fitting item was "Completeness", with respectively 21% and 28% more randomness in in-lying and outlying responses, comfortably within the requirements of effective measurement. "Body" was extremely overfitting, meaning that this item was highly predictive of overall results, and thus somewhat redundant, contributing less independent information than other items. Overall, however, the data sufficiently approximates the Rasch ideal of unidimensionality that the items are conducive to effective measurement.

Having established that teacher ratings provided an adequate measurement framework, the peer rating data was returned to the analysis, with the teacher raters anchored to maintain the previous measurement scale. Figure 2 compares the item difficulty estimates for the teacher ratings and the complete data set, with most items close to the linear trend line and 88% shared variance, suggesting that overall the peer raters and teacher raters were interpreting items similarly. However, item 6, "Speed", diverges somewhat from the linear trend, with students tending to assign higher ratings than teachers, suggesting that students may find fluency-related language features difficult to rate.

Figure 3 compares the estimates of person ability from teacher ratings and

Table 3 Item Measurement Report

Total Score	Total Count	Obs Ave	Fair-M Ave	Measure	Model *SE*	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr PtMea	Corr PtExp	Nu Items
389	213	1.8	2.06	-0.14	0.11	0.85	-1.7	0.90	-1.1	1.15	.68	.64	1 Confidence
298	212	1.4	1.58	1.04	0.12	1.03	0.3	1.01	0.1	0.97	.63	.62	2 Notes/reading
341	208	1.6	1.86	0.34	0.12	1.01	0.1	0.99	-0.1	1.01	.70	.63	3 Eyes
244	211	1.2	1.32	1.75	0.12	1.22	2.1	1.18	1.7	0.82	.58	.59	4 Hands
252	211	1.2	1.36	1.63	0.12	0.62	-4.4	0.64	-4.1	1.38	.66	.60	5 Body
376	211	1.8	2.02	-0.04	0.11	0.89	-1.2	0.91	-1.0	1.10	.66	.63	6 Speed
449	210	2.1	2.40	-1.05	0.12	1.10	1.0	1.14	1.3	0.80	.59	.64	7 Volume
314	209	1.5	1.71	0.72	0.12	0.76	-2.7	0.76	-2.7	1.26	.67	.62	8 Pausing
288	209	1.4	1.56	1.09	0.12	0.80	-2.1	0.80	-2.1	1.22	.63	.61	9 Intonation
505	211	2.4	2.65	-1.85	0.13	1.04	0.4	1.02	0.2	0.98	.63	.63	10 Organization
530	211	2.5	2.75	-2.29	0.14	1.06	0.6	0.98	-0.1	0.97	.61	.61	11 Relevance
461	210	2.2	2.46	-1.21	0.12	1.21	2.1	1.28	2.5	0.71	.62	.64	12 Completeness
370.6	210.5	1.8	1.98	0.00	0.12	0.97	-0.5	0.97	-0.4		.64		Mean (Count:12)
93.3	1.3	0.4	0.47	1.29	0.01	0.18	1.9	0.17	1.8		.03		*SD* (Pop.)
97.4	1.4	0.5	0.50	1.34	0.01	0.18	2.0	0.18	1.9		.04		*SD* (Sample)

Model, Popuin: RMSE .12 Adj (True) *SD* 1.28 Separation 10.68 Strata 14.57 Reliability .99
Model, Sample: RMSE .12 Adj (True) *SD* 1.34 Separation 11.16 Strata 15.21 Reliability .99
Model, Fixed (all same) chi-square: 1276.3 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.9 d.f.: 10 significance (probability): .36

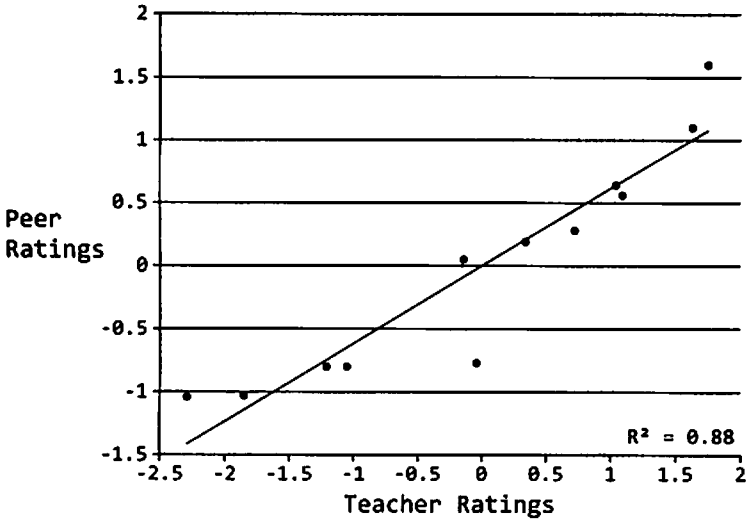

Figure 2 Comparison of item difficulty from teacher ratings and combined teacher and peer ratings. The relative difficulty of most items are similar, with the exception of item 6, "Speed".

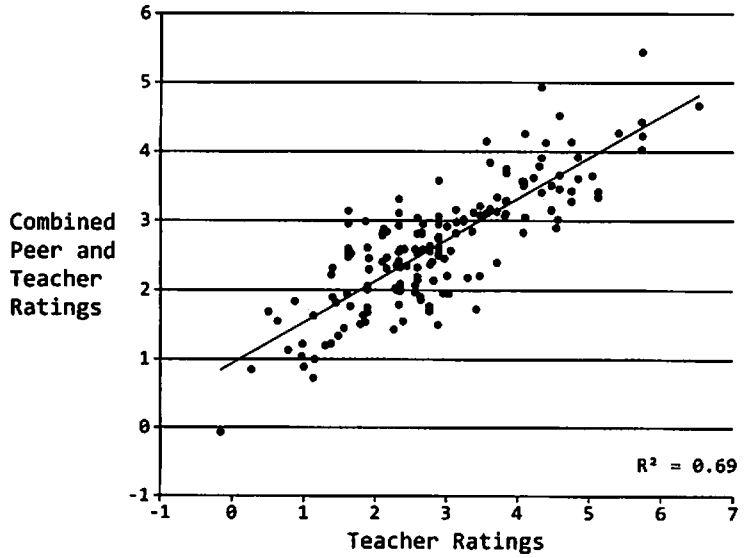

Figure 3 Person ability estimates from teacher raters and combined ratings by teachers and peer raters. Teacher ratings and peer ratings follow a linear trendline with 69% shared variance.

combined teacher and peer ratings. A very clear linear relationship is evident, with a moderate to strong raw correlation of .83 giving 69% shared variance. However, this correlation is attenuated by measurement error, with a reliability coefficient of .86 reported for the teacher ratings and .97 for the combined ratings, resulting in a disattenuated correlation of .91. Thus, for low stakes classroom assessment purposes, the teacher ratings and combined ratings are interchangeable, but the much greater number of observations available from the peer ratings greatly reduces measurement error, hence the very high reliability coefficient for the combined ratings.

Once it was established that peer assessment provides useful measurement for classroom purposes and that the items were usefully unidimensional, the data could be used to inform classroom instruction. Figure 1 mapped students, raters, and items on a common scale of measurement, making it obvious that the non-verbal communication items were posing the greatest difficulty for students, so preparation for the second presentation focused on practicing these items. The second presentation was graded in the same way as the first, using paper rating sheets in class and data entry assigned as homework, but the rubric was revised as previously shown in Table 1. In order to directly compare the first and second presentations, the three content items from the first presentation were averaged to give a single rating, while item 1, "Confidence", was deleted. The resulting longitudinal dataset adds a fourth facet, "Time", with the probability of success expected to increase in the second presentation. Figure 4 shows the resulting measurement rulers for teacher ratings. The "Time" column compares the difficulty of the first and second presentations. The first presentation was more difficult, meaning that higher ratings were assigned for the second presentation, evidence that the presentation instruction was effective. The increase in ability was 0.76 logits, meaning that a probability of success of 50% increased to 68%, a substantively large improvement.

Table 4 shows the measurement report comparing the first and second presentations. The logit measures increased from -0.38 to 0.38, representing a mean raw score increase from 1.6 to 1.9. After adjustment for raters, the fair-measure average shows an increase in raw score from 1.87 to 2.17. The reliability

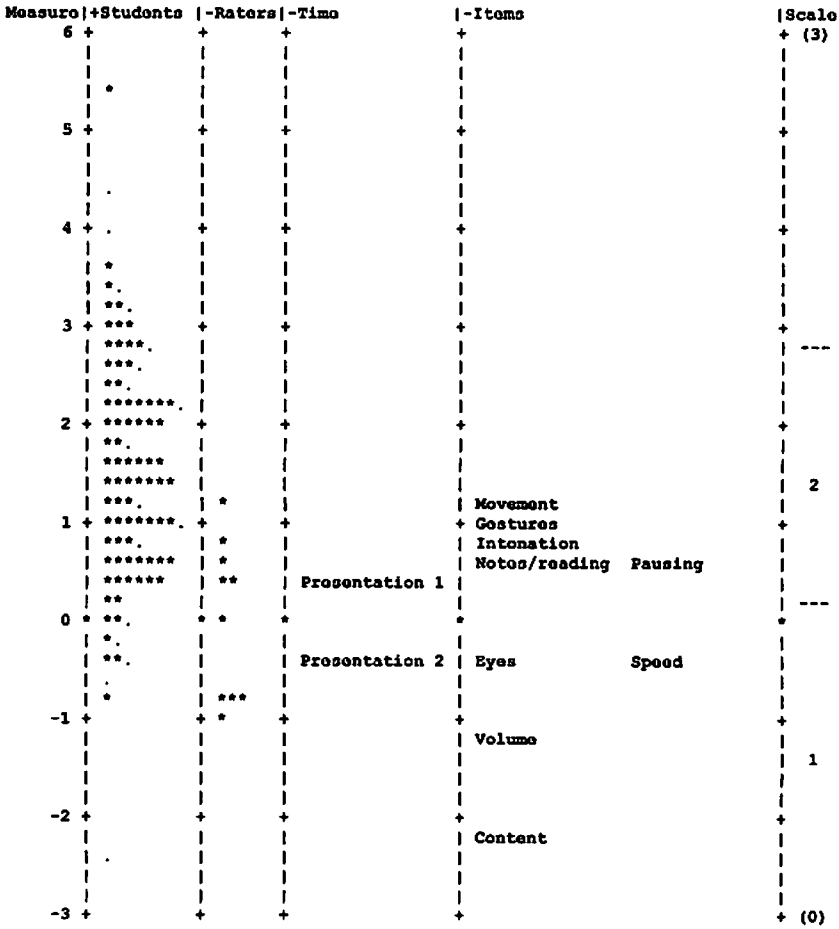

Figure 4 Facets measurement rulers for longitudinal results. Student ability, rater severity, time of administration, and item difficulty are mapped onto a shared measurement scale.

coefficient of .99 indicates very high confidence that the reported increase was not due to chance. The infit and outfit statistics are very close to the expected value of 1.00, with the second presentation being slightly more consistent than the first. These statistics indicate very good model-data fit, meaning that effective measurement was possible.

While Table 4 clearly shows overall growth, this is an average figure, but diagnostic uses require a more detailed analysis. Figure 5 shows how the relative difficulty of individual items changed between the first and second presentations. Although all items were relatively less difficult in the second presentation, the non-verbal communication items that were targeted for practice showed much larger improvements. This shows that students benefitted from instruction, providing compelling evidence of the value of MFRM for classroom diagnosis.

Table 5 shows the measurement report for teacher raters for the longitudinal data, with good overall model-data fit indicated by the infit and outfit mean-squared figures of 0.99 and standard deviations of 0.16. Rater 1, the author, again rated the second presentations three times, the first rating being of live presentations in class, reported as "Rater 1", with subsequent ratings reported as "Rater 1D" and "Rater 1E". Again, Rater 6 shows the highest levels of misfit, but the respective infit and outfit statistics of 1.29 and 1.31 are not substantively large enough to harm measurement, so this rater is performing consistently enough that grades between classes could be considered comparable, meaning that school wide assessment is possible.

Table 6 summarizes the rater fit statistics for the combined teacher and peer raters from the longitudinal dataset, and it is apparent that the peer raters are

Table 4 Time Measurement Report

Total Score	Total Count	Obs Ave	Fair-M Ave	Measure	Model *SE*	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr PtMea	Corr PtExp	Time
2964	1812	1.6	1.87	0.38	0.04	1.02	0.6	1.01	0.4	0.98	.71	.69	Presentation 1
3677	1890	1.9	2.17	-0.38	0.04	0.97	-0.9	1.00	0.0	1.02	.63	.66	Presentation 2
3320.5	1851.0	1.8	2.02	0.00	0.04	1.00	-0.1	1.01	0.3		.67		Mean
356.5	39.0	0.2	0.15	0.38	0.00	0.02	0.8	0.01	0.2		.04		*SD* (Pop)
504.2	55.2	0.2	0.21	0.54	0.00	0.04	1.1	0.01	0.2		.05		*SD* (Sample)

Model, Populn: RMSE .04 Adj (True) *SD* .38 Separation 9.50 Strata 13.01 Reliability .99
Model, Sample: RMSE .04 Adj (True) *SD* .53 Separation 13.48 Strata 18.31 Reliability .99
Model, Fixed (all same) chi-square: 182.7 d.f.: 1 significance (probability): .00

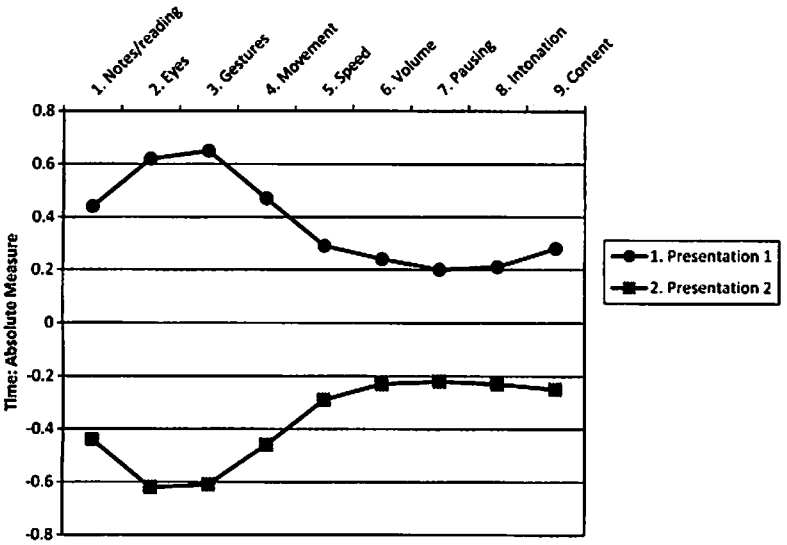

Figure 5 Interaction between items and time of presentation. The relative difficulties of items are shown for Presentation 1 and Presentation 2, with the size of improvement on each item shown by the vertical gap.

Table 5 Teacher Raters Measurement Report for Longitudinal Data

Total Score	Total Count	Obs Ave	Fair-M Ave	Meas	Model *SE*	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim Discr	Corr PtMea	Corr PtExp	Exact Obs %	Agree. Exp %	Raters
1056	528	2.0	1.85	0.42	0.07	1.29	4.5	1.31	4.3	0.62	.59	.68	48.9	47.2	Rater 6
902	476	1.9	2.37	−0.89	0.08	1.20	3.1	1.16	2.4	0.80	.70	.65	52.2	43.1	Rater 1
438	216	2.0	2.30	−0.71	0.11	1.14	1.4	1.08	0.7	0.94	.66	.61	56.3	45.3	Rater 1E
1037	514	2.0	1.74	0.69	0.08	0.93	−1.1	1.06	0.7	1.00	.63	.66	49.5	44.6	Rater 5
186	107	1.7	2.32	−0.76	0.16	0.99	0.0	0.95	−0.3	1.05	.74	.62	45.5	43.5	Rater 1B
405	225	1.8	2.06	−0.09	0.11	0.98	−0.1	0.96	−0.4	1.07	.68	.62	55.4	45.0	Rater 1D
914	639	1.4	1.56	1.16	0.07	0.96	−0.7	0.95	−0.8	1.04	.69	.69	44.1	40.3	Rater 3
31	18	1.7	2.41	−1.01	0.39	0.93	−0.1	0.86	−0.3	1.26	.85	.63	50.9	40.4	Rater 1C
1028	574	1.8	1.67	0.88	0.07	0.79	−3.9	0.80	−3.6	1.24	.71	.67	43.7	42.9	Rater 4
644	405	1.6	1.69	0.32	0.08	0.73	−4.2	0.76	−3.6	1.23	.64	.67	71.1	53.9	Rater 2
664.1	370.2	1.8	2.02	0.00	0.12	0.99	−0.1	0.99	−0.1		.69				Mean (Count: 10)
359.5	202.5	0.2	0.30	0.76	0.09	0.16	2.6	0.16	2.3		.07				*SD* (Pop)
378.9	213.4	0.2	0.32	0.80	0.10	0.17	2.7	0.17	2.4		.07				*SD* (Sample)

Model, Pop: RMSE .16 Adj (True) *SD* .75 Separation 4.80 Strata 6.74 Reliability (not inter-rater) .96
Model, Samp:RMSE .16 Adj (True) *SD* .79 Separation 5.07 Strata 7.10 Reliability (not inter-rater) .96
Model, Fixed (all same) chi-square: 637.3 d.f.: 9 significance (probability): .00
Model, Random (normal) chi-square: 8.6 d.f.: 8 significance (probability): .38
Inter-Rater agreement opportunities: 1667 Exact agreements: 849 = 50.9% Expected: 727.2 = 43.6%

performing much less consistently than the teacher raters. The respective mean infit and outfit values of 1.17 and 1.31, and standard deviations of 0.32 and 0.49 indicate that many peer raters are highly misfitting, meaning that they are interpreting the rubric quite differently from teachers.

Detailed diagnosis of problematic ratings is available from *Facets* in the form of a table of unexpected responses, a small sample of which is shown in Table 7. Unexpected responses are calculated by comparing the observed score actually assigned by a rater to a performance on an item by a student, with the statistically predicted score, the difference between these being termed the "residual". For example, the first line in Table 7 shows that Rater 436 gave a score of 2 to Student 422 for item 9 in the second presentation, but the predicted score was 3.0. This resulted in a score residual of -1.0 and a standardized residual of -9.0, meaning that the difference is 9.0 standard deviations different from the expected result. Table 7 only shows the 20 most unexpected responses for brevity, but *Facets* reports standardized residuals greater than 3.0 by default, a setting that can be adjusted according to the needs of the situation.

The table of unexpected responses is an invaluable resource for diagnostic purposes as it allows teachers to identify and address specific difficulties by individual students. Looking at Table 7, it is obvious that most difficulties arose from item 9 "Content", an item so problematic that all students would benefit from explicit instruction on and practice of it. The largest raw residual in Table 7 is -2.0, where Rater 523 rated Student 513 on item 6, "Volume". Examination of the complete table of unexpected responses will reveal whether this was simply an isolated problem, perhaps just a mistake in data entry, or part of a larger pattern of

Table 6 Combined Teacher and Peer Raters Summary Statistics

Total Score	Total Count	Obsvd Average	Fair-M Average	Measure	Model *SE*	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Point-Measure Correlation	
563.7	246.5	2.3	2.42	-1.13	0.13	1.17	1.5	1.31	2.0	.44	Mean (Count: 185)
173.2	87.2	0.3	0.30	0.95	0.04	0.32	2.9	0.49	2.8	.14	*SD* (Population)
173.7	87.4	0.3	0.30	0.95	0.04	0.32	2.9	0.49	2.9	.14	*SD* (Sample)

Model, Pop: RMSE .13 Adj (True) *SD* .94 Separation 7.07 Strata 9.76 Reliability (not inter-rater) .98
Model, Samp: RMSE .13 Adj (True) *SD* .94 Separation 7.09 Strata 9.79 Reliability (not inter-rater) .98
Model, Fixed (all same) chi-square: 12640.1 d.f.: 184 significance (probability): .00
Model, Random (normal) chi-square: 180.2 d.f.: 183 significance (probability): .55
Inter-Rater agreement opportunities: 367148 Exact agreements: 171719 = 46.8% Expected: 171454.5 = 46.7%

Table 7 Unexpected Responses

Observed Score	Expected Score	Score Residual	Standardized Residual	Student Number	Rater Number	Time	Item
2	3.0	-1.0	-9.0	422	436	Presentation 2	9 Content
2	3.0	-1.0	-9.0	423	436	Presentation 2	9 Content
2	3.0	-1.0	-9.0	433	436	Presentation 2	6 Volume
1	3.0	-2.0	-9.0	513	523	Presentation 1	6 Volume
2	3.0	-1.0	-9.0	519	525	Presentation 2	9 Content
2	3.0	-1.0	-9.0	562	564	Presentation 2	9 Content
2	3.0	-1.0	-9.0	563	561	Presentation 2	6 Volume
2	3.0	-1.0	-9.0	563	564	Presentation 2	9 Content
2	3.0	-1.0	-8.6	448	451	Presentation 2	5 Speed
2	3.0	-1.0	-8.2	428	436	Presentation 2	9 Content
2	3.0	-1.0	-8.2	608	604	Presentation 2	9 Content
2	3.0	-1.0	-8.1	501	508	Presentation 2	9 Content
2	3.0	-1.0	-8.1	504	506	Presentation 2	9 Content
2	3.0	-1.0	-8.0	501	497	Presentation 2	9 Content
2	3.0	-1.0	-8.0	608	602	Presentation 2	9 Content
2	3.0	-1.0	-7.9	562	557	Presentation 2	9 Content
2	3.0	-1.0	-7.8	614	604	Presentation 2	9 Content
2	3.0	-1.0	-7.7	547	548	Presentation 2	9 Content
2	3.0	-1.0	-7.7	571	560	Presentation 1	9 Content
2	3.0	-1.0	-7.6	571	566	Presentation 1	9 Content

difficulties by this rater. The rater can then be asked to re-rate problematic items using video recorded performances in order to gain more practice or to diagnose the cause of the problems. In such a manner, *Facets* analysis of peer assessment can provide very fine-grained diagnostic feedback to guide classroom instruction.

Conclusions and Future Directions

The major objective of this preliminary investigation was to test the MOARS peer assessment module under operational conditions to confirm that it could provide a practical solution to peer assessment data collection. This objective was met in full and encountered no significant problems, MOARS proving to be simple to use and able to output data for immediate analysis.

Having established the practicality of peer assessment, a series of questions needed to be resolved in order to demonstrate the validity of basing classroom decisions on peer assessment. The first requirement was to investigate the performance of the teacher raters, a prerequisite for any further analysis. Although teachers varied in the severity of their ratings, as previous research has repeatedly reported, the data-model fit for raters was sufficiently good that all students could be compared on a shared scale for low stakes purposes. However,

data-model fit improved for the second presentation, supporting the importance of rater training. Improved training videos and more extensive rater training need to be addressed in future studies.

A second prerequisite for effective measurement was adequate item performance. Data-model fit for items was found to sufficiently approximate the assumptions of the Rasch model that effective measurement was possible, and the relative difficulties of items when rated by teachers and students were very similar, with the exception of one item, "Speed". This supports the argument that peer ratings can provide useful information to inform low stakes classroom decisions. However, the data-model fit of the peer raters was much worse than that of the teacher raters, so, rather than viewing peer assessment as a tool for measuring proficiency, it is better viewed as a diagnostic tool to identify problematic patterns of responses, allowing raters with idiosyncratic interpretations of the rubric to be identified.

This demonstration that the requirements of effective measurement were satisfied allowed proficiency growth to be measured, demonstrating the effectiveness of instruction. Although performance on all rubric items showed substantive improvement, the non-verbal communication items that were specifically targeted for instruction showed substantively larger gains. This provides solid evidence of the effectiveness of instruction and a compelling illustration of the value of MFRM in program evaluation.

Although this proof-of-concept study achieved its objectives, considerable work remains to be done to improve the practicality of MFRM as a classroom diagnostic tool. *Facets* provides extremely detailed summaries of students, items, and raters, but novice users are overwhelmed by the volume of information provided in the output tables, so simplified graphical summaries are an essential next step. The current MOARS package provides simple graphical summaries of raw score results, so future efforts will focus on providing analogous graphical summaries of the diagnostic outputs from *Facets* for use in the 2012 academic year.

References

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

Associates.

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.

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 28 March, 2008, from http://tinyurl.com/zocgc

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

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.

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

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

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

Negishi, J. (2010). Multi-faceted Rasch analysis for the assessment of group oral interaction using CEFR criteria. *Annual Review of English Language Education in Japan, 21*, 111-120.

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

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

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

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology: An International Journal of Experimental Educational Psychology, 25*(6), 631 - 645.

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

Appendix 1 Grading Rubric

Presentation Reviews
Please watch other students making presentations. Rate each presentation 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 を塗りつぶして下さい。
"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. Non-verbal Communication(非言語コミュニケーション)
- **Confidence.** Does the presenter look relaxed and confident?
  自信:リラックスして、自信をもって発表しているように見えますか。
- **Notes/Reading.** Does the presenter use notes to remember key points without just reading continuously?
  メモ / 読み上げ: 書いてあることを全て棒読みするのではなく、必要なことのみを書いたメモを適切に使えていますか。
- **Eyes.** Does the presenter look at the audience regularly?
  アイコンタクト:定期的に聞き手の方を見ながら発表できていますか。
- **Hands.** Does the presenter use their hands naturally?
  手の動き:自然な手の動きでジェスチャーが使えていますか。
- **Body.** Does the presenter move naturally, not frozen like a statue, but not moving too much?
  身体の動き:銅像のように全く動かなかったり、動き過ぎたりせず、自然な動きですか。

2. Voice(声)
- **Speed.** Does the presenter speak at a natural speed that is easy to understand?
  スピード:理解しやすい、自然なスピードで話せていますか。
- **Volume.** Does the presenter speak loudly enough to listen to easily?
  声の大きさ:聞きやすい声の大きさで話せていますか。
- **Pausing.** Does the presenter pause naturally when they speak?
  間の取り方:話しながら自然なところで一息置いたりといった間が取れていますか。
- **Intonation.** Does the presenter vary their intonation naturally?
  イントネーション:質問で語尾を上げたり、重要なところを強く言ったりなどのイントネーションが使えていますか。

3. Contents and Organization(内容と構成)
- **Organization.** Is the information organized in a logical way that is easy to understand?
  構成:発表の内容が、わかりやすい、筋の通った構成になっていますか。
- **Relevance.** Is the information relevant to the presenter's key points?
  関連性:発表者の言いたいこと(論点)に関連した情報が含まれていますか。
- **Completeness.** Is there enough information to completely understand the presenter's key points?
  完全性:発表者の言いたいこと(論点)が完全に伝わるのに十分な量の情報が含まれていますか。