教師による評価と学生による相互評価の実践

Teacher Assessment and Peer Assessment in Practice

中村 優治 NAKAMURA, Yuji
● 東京経済大学
Tokyo Keizai University

Keywords 代替評価，ラッシュモデル，学生相互評価，オーラルプレゼンテーション
Alternative Assessment, Rasch Model, Peer Assessment, Oral Presentation

ABSTRACT

本稿はオーラルプレゼンテーションクラスの評価方法として，教員による評価と学生による相互評価の両方を取り入れ，ラッシュモデルの FACET 理論を適用して項目応答理論によるデータ分析を行い，より科学的で実践的な評価方法の提言を試みたものである．結論として次の 3 点が検証された．1) 学生相互評価は勉学への動機付けとなりうる．2) 学生はある一定の信頼性を備えた評価者となりうる．3) ラッシュモデル（FACET 理論）により評価者の評価特性，学生の能力の特徴，および評価項目の適切さが明確に示され，プレゼンテーション能力の科学的な説明が可能となる．

1. Introduction

Alternative assessment as classroom-based language assessment has become common in classroom settings, while computer-based language testing as a formal assessment has made great progress in the field of education.

This paper attempts to address some issues facing alternative assessment, with particular reference to teacher assessment and peer assessment of oral presentation skills in English in classroom settings in Japan.

The field of self-assessment of language proficiency is concerned with questions of how, under what conditions, and with what effects learners and other users of a foreign or second language may judge their own ability in the language (Oscarson 1997). He further claims that techniques and materials used for the purpose of self-assessment can include self-reports, self-testing and mutual peer-assessment.

In the present research, mutual peer assessment is focused on to investigate possible ways of realizing the goal of the oral presentation class; how we can enhance the students' communication ability by involving them in in-class activities (presentation and assessment).

The present paper also explores what the Rasch measurement model, which is powerful in handling polytomous data involving raters' judgment (Linacre 1989, 1994), analyzing the multi-faceted data of peer assessment in the presentation class. This is because the theoretical and statistical support for the peer assessment has not been well established. We look at the data collected from the actual oral presentation course in a university classroom setting. The data will be analyzed from the viewpoint of not only language testing and linguistic aspects but also speech communication.

Practical advantages as well as problems involved in peer assessment and teacher assessment will be discussed.

We focus on the utility and importance of the alternative assessment by highlighting the classroom situations where teachers and students cordially work together to attain the goal of the presentation course.

2. Context

Communication skills are a highly desirable aspect in today's job market and increasingly rapid changes in the workplace make management aware of the importance of competent communicators (Tatum 1997a). It follows that as business grows on an increasingly global level, students are in need of English oral communication skills as they graduate from university if they are to be competitive in the job market. Communication classes are now firmly entrenched in universities that teach English as a second or foreign language. However, many students are still graduating with little more than elementary "survival English" skills.

In addressing the question of what are the appropriate expectations of proficiency of university students, our initial assumptions are based on a study of Japanese graduate students at a university in Japan. In the study where Japanese graduate students were asked to cite what they felt were the most important / useful English skills for them to learn, the resulting list was conversation, presentation, discussion, and debate (Hiyoshi Review,

2000). The oral presentation skill is the area we decided to focus on.

The present program takes place in a small class of a university where 12 students take a whole year oral presentation course. They are highly motivated to learn oral presentation skills because they chose this course at the beginning of the academic year among many choices of courses. The teacher is a Japanese teacher of English majoring in Applied Linguistics and Language Testing. Also, he has been trained in an oral presentation course in the United States.

3. Description

In class, twelve students gave public speaking presentations that were assessed by five raters (one classroom teacher and four students who were chosen at random) using eleven evaluation items (e.g. sincerity, eye contact, and oral fluency). Three facets of students' ability, item difficulty, and raters' severity plus rating categories will be thoroughly discussed.

1) Subjects
   Twelve university students
2) Raters
   Five raters (one teacher and four students who are chosen at random from among the twelve students above)
3) Rating items
   Tatum's items (1997b) were partly used and arranged for this present research.
   1. speaker's sincerity to the audience
   2. oral fluency
   3. pronunciation (sonority or enunciation)
   4. eye contact
   5. facial expression
   6. appropriate language (grammar)
   7. originality of expressions
   8. content (target of the unit)
   9. written fluency (smooth flow of speech)
   10. appropriate evidence
   11. holistic evaluation (overall impression)
4) Rating scale
   Items 1 through 10 were rated on a six-point scale (1 is poor and 6 is good), while only item 11 was judged on a four-point scale (1 is poor and 4 is good).

These twelve students took turns to give presentations and while one student was giving his / her presentation, the others evaluated the speaker's presentation using the scoring criteria mentioned above.

After the students gave presentations and finished scoring, several steps were taken. First, the teacher collected the scores from the students and calculated them and decided the top three students who got the higher points. Then, all the students shared their opinions and thoughts on the good points of the top three students so that for the next time those comments will be good guidelines for others' better presentations. Finally, the teacher gave summary comments not only on the top three students but also on the whole class in terms of the class target, linguistic aspects, language testing or assessment and better presentation skills.

4. Distinguishing Features

The uniqueness of this case study is the use of a new statistical program called the Many-Facet Rasch Measurement Model to investigate three facets (the raters, students, items) on the same continuum (scale) and to improve

the test itself by deleting misfitting factors. The procedure is as follows:

The data was analyzed using the Many-Facet Rasch Measurement Model, which was able to give detailed information about three facets of the study (student ability, item difficulty, and rater severity). The data were investigated mainly from the viewpoint of unexpected scores and fit statistics. Also, a benchmark of the acceptable range of the infit and outfit statistics was set between 0.6-1.4 by taking into consideration that this is a performance speech test data which involves rater's judgment. Furthermore, Separation index for the students measurement report should be over 2.0 in theory.

4. 1. What does the Many-Facet Rasch Measurement tell us about the test facets (students, items, raters) and rating categories ?

First let us look at the unexpected responses in Table 1.

Table 1 shows three unexpected responses. In the first case, rater 1, student 12 and item 11 are interrelated to result in score 2, whose expected one is 3.7. Then, in the second case, rater 2, student 5 and item 9 are interrelated to produce score 6 whose expected one is 3.7. Furthermore, in the third case, rater 2, student 6 and item 5 functioned with each other and ended up with the score with 6 whose expected one is 3.8. Although it is not as easy to tell what is the cause of the discrepancy between the observed scores and the expected scores in these cases, rater 2 may have something to do with this phenomenon because of its frequent appearance in this unexpected data. Thus, this table of unexpected responses can lead us to a further investigation of the noticeable facets.

Now let us examine the raters' measurement in Table 2.

Table 2 indicates the raters measurement report. According to our benchmark of the fit statistic about the acceptable range (0.6-1.4) for this research where raters' judgment is involved in a speaking performance test, all the raters are working rather reasonably except rater 2 whose infit statistic is 1.5, which is beyond the maximum range (1.4). When we look at the measure column, rater 1 (the teacher) is the most lenient followed by rater 4, while rater 2 is the severest among the five raters. We also notice, as mentioned jut now, that the teacher is more lenient than the students raters.

We noted also is that Separation index 6.18 is a little big, which means that the raters' judgments vary greatly. However, these students raters as a whole do good jobs as raters with each one's level of severity rather consistently, which was shown in the fit statistic.

Table 1. Unexpected Responses

Cat	Step	Exp.	Resd	StRes	Nr	Nu	st	Nu	items
2	2	3.7	-1.7	-3	1	12	12	11	holistic evaluation
6	6	3.7	2.3	3	2	5	5	9	wtitten fluency
6	6	3.8	2.2	3	2	6	6	5	facial expression
Cat	Step	Exp.	Resd	StRes	Nr	Nu	st	Nu	items

One thing we should pay attention to is the cause of the misfit of rater 2. It might be difficult to relate this severity and the misfit result; however, it is worth investigating the reason because rater 2 is highly involved in two of three unexpected responses as shown above in Table 1.

Let us go on to the students' measurement in Table 3. Table 3 presents the students measurement report. In other words, it shows students' ability. The measure column indicates that student 8 is the most able

Table 2. raters Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N raters
611	132	4.6	4.65	1.05	.11	1.3	2	1.3	2	1 1
459	132	3.5	3.47	-.93	.13	1.5	3	1.4	2	2 2
498	132	3.8	3.75	-.36	.12	.6	-3	.6	-3	3 3
576	132	4.4	4.36	.64	.11	.8	-1	.9	-1	4 4
495	132	3.8	3.73	-.40	.12	.7	-2	.7	-2	5 5
527.8	132.0	4.0	3.99	.00	.12	1.0	-.4	1.0	-.4	Mean (Count: 5)
56.5	.0	.4	.44	.73	.01	.3	2.7	.3	2.7	S.D.

RMSE (Model) .12 Adj S.D. .72 Separation 6.18 Reliability .97
Fixed (all same) chi-square: 200.0 d.f.: 4 significance: .00
Random (normal) chi-square: 4.0 d.f.: 3 significance: .26

Table 3. students Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu students
221	55	4.0	3.98	.57	.18	1.3	1	1.4	1	1 1
205	55	3.7	3.69	.05	.19	.7	-1	.7	-1	2 2
217	55	3.9	3.91	.44	.18	.9	0	.9	0	3 3
196	55	3.6	3.54	-.27	.19	.8	-1	.8	-1	4 4
227	55	4.1	4.10	.75	.18	1.1	0	1.1	0	5 5
243	55	4.4	4.41	1.23	.17	1.6	2	1.7	3	6 6
218	55	4.0	3.93	.47	.18	1.0	0	.9	0	7 7
246	55	4.5	4.48	1.32	.17	.6	-2	.6	-2	8 8
227	55	4.1	4.10	.75	.18	.8	-1	.8	-1	9 9
197	55	3.6	3.55	-.23	.19	1.2	0	1.1	0	10 10
206	55	3.7	3.71	.08	.18	.8	0	.8	-1	11 11
236	55	4.3	4.27	1.02	.17	.9	0	1.0	0	12 12
219.9	55.0	4.0	3.97	.51	.18	1.0	-.2	1.0	-.2	Mean (Count: 12)
16.1	.0	.3	.30	.51	.01	.3	1.4	.3	1.5	S.D.

RMSE (Model) .18 Adj S.D. .48 Separation 2.66 Reliability .88
Fixed (all same) chi-square: 95.5 d.f.: 11 significance: .00
Random (normal) chi-square: 11.0 d.f.: 10 significance: .36

followed by student 6, whereas student 4 is the poorest. The fit statistics show that all the students fit the model except student 6 whose infit and outfit statistic scores are over the acceptable range (0.6-1.4). Whether the high ability of this student 6 is related to the misfit result is not clear because student 8 whose ability is the highest did not influence any unexpected responses in Table. However, it could be said that student 6 could have affected the unexpected score with relation to the other facets in the third case of item 5 in Table 1.

The separation index 2.66 of this measure is acceptable as an indicator of separating students because it meets the requirements of the acceptable score 2.0. It can be said that this presentation test was able to separate the students reasonably.

Next let us investigate the items' measurement in Table 4.

Table 4 indicates the items measurement report. The fit statistics prove that all the items are functioning well within the acceptable range. It can be said that on the whole, all the items fit the model. The measure column suggests that the easiest item is item 11 (holistic evaluation) while the hardest items are item 5 (facial expression) and item 7 (originality). One interpretation for item 11 is that a 4-point scale is used only for this item (holistic evaluation) so raters' judgment did not spread widely within this scale. Another interpretation for item 5 (facial expression) is that students are not as familiar with facial expressions even in Japanese conversations because of the classroom culture. Still another interpretation for item 7 (originality) is that since students are assigned to use the target of the chapter their choice was limited to expand their ideas more freely even though they were allowed to choose their own topics. Furthermore, students tend not to stand out among peers in

Table 4. **items Measurement Report**

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu items
256	60	4.3	4.22	.22	.17	.9	0	.8	0	1 sincerity
250	60	4.2	4.11	.05	.17	.8	0	.9	0	2 oral fluency
256	60	4.3	4.22	.22	.17	.7	-2	.7	-1	3 pronunciation
257	60	4.3	4.23	.24	.17	.9	0	.9	0	4 eye contact
236	60	3.9	3.88	-.36	.17	1.2	1	1.2	1	5 facial expression
243	60	4.1	4.00	-.15	.17	1.2	1	1.2	1	6 grammar
236	60	3.9	3.88	-.36	.17	.7	-1	.7	-1	7 originality
244	60	4.1	4.01	-.12	.17	1.0	0	.9	0	8 content
246	60	4.1	4.04	-.07	.17	.9	0	.9	0	9 wtitten fluency
235	60	3.9	3.87	-.39	.17	1.1	0	1.2	0	10 evidence
180	60	3.0	3.02	.72	.19	1.3	1	1.4	1	11 holistic evaluation
239.9	60.0	4.0	3.95	.00	.17	1.0	-.1	1.0	-.2	Mean (Count: 11)
20.5	.0	.3	.32	.32	.01	.2	1.2	.2	1.2	S.D.

RMSE (Model) .17 Adj S.D. .27 Separation 1.56 Reliability .71
Fixed (all same) chi-square: 34.6 d.f.: 10 significance: .00
Random (normal) chi-square: 9.7 d.f.: 9 significance: .37

class by doing extremely original things.

Let us look at the separation index of 1.56, which is below 2.0 (a suggested point initially). It may be that all the items do not necessarily function well to spread the students on the scale. This is probably because the number of items is not enough to separate the students' various abilities, so that some extremely good students or extremely poor students were not well measured by these items.

Let us look at All Facet Vertical Rulers in Table 4'. It is clear that the columns of students and items in Table 4' provide us with this fact. The students are more spreading on the scale while items are not spreading as widely as the students. On the whole, however, all the items function well to measure these 12 students.

Table 4'. **All Facet Vertical Rulers**

Measr	+raters	+students	+items			S.1	S.2
2						(6)	(4)
						5	
		8					
		6				—	3
1	1	12					
		5 9					
			holistic evaluation				
	4	1					
		7					
		3				4	
			eye contact	pronunciation	sincerity		—
		11					
*0	*	*2	*oral fluency				
			content	wtitten fluency			
		10	grammar				
		4					
	3 5		evidence	facial expression	originality		
						—	
	2						
-1						(2)	(1)
Measr	+raters	+students	+items			S.1	S.2

Let us take a look at the functioning of the rating items (1-10) in Table 5a. Tables 5a shows the category (rating items) statistics. Items 1-10 were rated on a 1-6 point scale, although scale 1 was not used at all. The scores in the outfit column indicates that all the scales (2-6) were reasonably used and there were no misfitting scales among them. One thing that we should pay attention to is that the lowest category (rating item) is never used by these raters. This is a typical human action especially in a classroom situation. Peer students and teachers tend to avoid the lowest scale because they do not want to hurt others by giving them disappointing scores even if the raters are not mentioned. Therefore, we still need this unused bottom scale for the sake of students and teachers in a classroom setting.

Then let us also look at the functioning of the rating item (11) in Table 5b. Table 5b presents another category statistic for item 11 which was rated on a 1-4 point scale. Although four different categories were used, category 2 showed a big misfit as seen in the outfit statistic column. This is probably causing the unexpected response in Table 1, where item 11 was pointed out as an unexpected response, and category 2 was given to the one whose expected score was 3.7. It is not clear how the three facets (rater, student and item) are complicatedly interrelated with this category 2, but this category 2 has something to do with the unexpected score. In this way, we

Table 5a. **Category Statistics.**

Model = ?, ?, 1-10, R6

DATA Category Score	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	STEP CALIBRATIONS Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	THURSTONE THRESHOLD at	Cat PEAK Prob
2	11	2%	2%	-.10	-.62	1.2			(-4.13)		low	low	100%
3	147	25%	26%	-.21*	-.19	1.0	-3.01	.31	-1.79	-3.18	-3.01	-3.08	63%
4	270	45%	71%	.33	.38	.9	-.52	.10	.38	-.61	-.52	-.56	56%
5	116	19%	91%	.98	.99	1.0	1.53	.11	1.86	1.16	1.53	1.28	38%
6	56	9%	100%	1.69	1.51	.8	2.00	.16	(3.39)	2.70	2.00	2.36	100%
									(Mean)		(Modal)	(Median)	

Table 5b. **Category Statistics.**

Model = ?, ?, 11, R4

DATA Category Score	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	STEP CALIBRATIONS Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	THURSTONE THRESHOLD at	Cat PEAK Prob
1	1	2%	2%	-.48	.25	.7			(-3.45)		low	low	100%
2	16	27%	28%	1.02	.65	1.8	-2.34	1.02	-.97	-2.46	-2.34	-2.39	66%
3	25	42%	70%	1.19	1.21	1.1	.47.	32	1.20	.22	.47	.33	50%
4	18	30%	100%	1.58	1.83	1.3	1.86	.32	(3.10)	2.27	1.86	2.04	100%
									(Mean)		(Modal)	(Median)	

can explore the relationship between the unexpected scores and the category statistic integratively.

4. 2. How can the assessment or the test be improved by taking into consideration the Rasch-based analyzed data ?

In order to improve the test statistically using the Rasch program, it is theoretically easy to delete the misfitting items, students, raters. And the remains will be regarded as the modified test item. However, the case of deleting raters is not as easy as that of deleting students and items because the number of raters is usually not big. Therefore, even when only one rater is deleted, the effect of deletion to the whole is huge. Accordingly, the deletion of raters should be the last method for this test improvement.

How can we improve the test ? To begin with, let us look back at Table 1 and examine the details. In that table, three cases (in which a rater, an item and a student are interrelated) are detected as unexpected responses. Now, let us delete the combinations of three facets in three cases: case one (rater1, student 12, item 11), case two (rater 2, student 5, item 9), and case three (rater 2, student 6, item 5). Tables 6, 7, and 8 below show the results. Table 6 indicates no misfitting rater in the column of infit and outfit statistic. Table 7 shows one misfitting student (student 6) in the column of infit and outfit statistic. Table 8 presents no misfitting item in the column of infit and outfit statistic.

From the information above, Table 7 points out that only student 6 is still misfitting. Although student 6 is misfitting, Table 1 has indicated that student 6 and rater 2 were not functioning well with each other in the unexpected response combination. It seems that rater 2 is complicatedly connected with this misfitting student. Therefore, we should take the last method by deleting rater 2 in the analysis.

Let us look at the reanalyzed data in Tables 9, 10 and 11. Table 9 gives a satisfactory result without rater 2. Table 10 shows all the fit students within the acceptable range between 0.6-1.4. However, Table 11 produces

Table 6. Raters Measurement Report

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N raters
609	131	4.6	4.67	1.12	.11	1.3	2	1.3	2	1 1
447	130	3.4	3.43	-1.03	.13	1.4	2	1.3	2	2 2
498	132	3.8	3.75	-.36	.12	.6	-3	.6	-3	3 3
576	132	4.4	4.36	.67	.11	.9	0	.9	0	4 4
495	132	3.8	3.73	-.40	.12	.7	-2	.7	-2	5 5
525.0	131.4	4.0	3.99	.00	.12	1.0	-.4	1.0	-.4	Mean (Count: 5)
58.9	.8	.4	.46	.78	.01	.3	2.5	.3	2.4	S.D.

RMSE (Model) .12 Adj S.D. .77 Separation 6.49 Reliability. 98
Fixed (all same) chi-square: 218.3 d.f.: 4 significance: .00
Random (normal) chi-square: 4.0 d.f.: 3 significance: .26

Table 7. **Students Measurement Report**

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu students
221	55	4.0	3.98	.57	.18	1.4	1	1.4	2	1 1
205	55	3.7	3.69	.03	.19	.8	-1	.7	-1	2 2
217	55	3.9	3.91	.44	.18	.9	0	1.0	0	3 3
196	55	3.6	3.54	-.29	.19	.8	0	.8	0	4 4
221	54	4.1	4.04	.69	.18	1.0	0	1.0	0	5 5
237	54	4.4	4.36	1.19	.18	1.5	2	1.6	2	6 6
218	55	4.0	3.93	.47	.18	1.0	0	.9	0	7 7
246	55	4.5	4.47	1.35	.18	.7	-2	.7	-2	8 8
227	55	4.1	4.09	.76	.18	.8	-1	.8	-1	9 9
197	55	3.6	3.55	-.25	.19	1.2	0	1	.1 0	10 10
206	55	3.7	3.71	.07	.19	.9	0	.8	-1	11 11
234	54	4.3	4.33	1.10	.18	.8	-1	.8	0	12 12
218.8	54.7	4.0	3.97	.51	.18	1.0	-.2	1.0	-.3	Mean(Count: 12)
15.1	.4	.3	.30	.52	.01	.3	1.3	.3	1.4	S.D.

RMSE (Model) .18 Adj S.D. .49 Separation 2.67 Reliability. 88
Fixed (all same) chi-square : 96.3 d.f.: 11 significance: .00
Random (normal) chi-square : 11.0 d.f.: 10 significance: .36

Table 8. **Items Measurement Report**

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu items
256	60	4.3	4.21	.23	.17	.9	0	.9	0	1 sincerity
250	60	4.2	4.11	.06	.17	.9	0	.9	0	2 oral fluency
256	60	4.3	4.21	.23	.17	.7	-1	.7	-1	3 pronunciation
257	60	4.3	4.23	.26	.17	1.0	0	.9	0	4 eye contact
230	59	3.9	3.84	-.43	.18	1.1	0	1.1	0	5 facial expression
243	60	4.1	3.99	-.15	.17	1.2	1	1.2	1	6 grammar
236	60	3.9	3.88	-.36	.18	.8	-1	.8	-1	7 originality
244	60	4.1	4.01	-.12	.17	1.0	0	.9	0	8 content
240	59	4.1	4.00	-.13	.17	.8	-1	.7	-1	9 wtitten fluency
235	60	3.9	3.87	-.39	.18	1.2	0	1.2	1	10 evidence
178	59	3.0	3.05	.78	.19	1.3	1	1.3	1	11 holistic evaluation
238.6	59.7	4.0	3.95	.00	.18	1.0	-.1	1.0	-.2	Mean (Count: 11)
21.1	.4	.3	.31	.34	.01	.2	1.1	.2	1.1	S.D.

RMSE (Model) .18 Adj S.D. .30 Separation 1.69 Reliability. 74
Fixed (all same) chi-square: 38.7 d.f.: 10 significance: .00
Random (normal) chi-square: 9.8 d.f.: 9 significance: .37

Table 9. **Raters Measurement Reportt**

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N raters
609	131	4.6	4.62	.95	.11	1.3	2	1.3	2	1 1
498	132	3.8	3.75	-.68	.13	.8	-1	.8	-1	3 3
576	132	4.4	4.32	.47	.12	.9	-1	.9	0	4 4
495	132	3.8	3.73	-.74	.13	.9	0	.9	-1	5 5
544.5	131.8	4.1	4.11	.00	.12	1.0	-.3	1.0	-.3	Mean (Count: 4)
49.4	.4	.4	.38	.73	.01	.2	1.7	.2	1.6	S.D.

RMSE (Model) .12 Adj S.D. .72 Separation 5.87 Reliability .97
Fixed (all same) chi-square : 143.4 d.f.: 3 significance: .00
Random (normal) chi-square : 3.0 d.f.: 2 significance: .22

Table 10. **Students Measurement Report**

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu students
179	44	4.1	4.02	.70	.21	1.3	1	1.3	1	1 1
168	44	3.8	3.79	.19	.22	.9	0	.8	0	2 2
183	44	4.2	4.10	.87	.21	.8	0	.8	0	3 3
160	44	3.6	3.61	-.22	.23	.9	0	.8	0	4 4
182	44	4.1	4.08	.83	.21	.9	0	.9	0	5 5
203	44	4.6	4.58	1.70	.20	1.3	1	1.3	1	6 6
184	44	4.2	4.13	.91	.21	1.0	0	1.0	0	7 7
202	44	4.6	4.56	1.66	.20	.7	-1	.7	-1	8 8
192	44	4.4	4.31	1.25	.20	.7	-1	.7	-1	9 9
165	44	3.8	3.72	.04	.23	1.4	1	1.3	1	10 10
171	44	3.9	3.85	.33	.22	1.0	0	1.0	0	11 11
189	43	4.4	4.34	1.28	.20	.7	-1	.8	-1	12 12
181.5	43.9	4.1	4.09	.79	.21	1.0	.2	1.0	-.3	Mean(Count: 12)
13.2	.3	.3	.30	.59	.01	.2	1.2	.2	1.0	S.D.

RMSE (Model) .21 Adj S.D. .55 Separation 2.61 Reliability.87
Fixed (all same) chi-square: 91.2 d.f.: 11 significance: .00
Random (normal) chi-square: 11.0 d.f.: 10 significance: .36

another misfitting item (item 3: pronunciation), whose infit and outfit statistic scores are below the acceptable range. What is left for us to either delete this item or not. Since "pronunciation" is an indispensable part of speaking skill, it might be wise of us to leave this item as is. Nevertheless, we need to have more detailed discussion about the content of "pronunciation" in the rater training session so that all the raters fully understand its crite-

rion for the better evaluation. Also, since rater 2 still exists in reality, there should be a training session or more discussion about the evaluation criteria from this viewpoint as well in order to improve the test and the rater in the classroom situation.

5. Practical Ideas

How does the peer assessment (including the teacher assessment) work in a classroom setting ? As shown in Table 2, all the raters (one teacher and four students) except rater 2 did very well in their evaluation by maintaining their leniency or severity, although the variance of harshness was rather wide. Also indicated in Table 4, they used 11 rating items well. It may be that students' raters understood the rating criterion for evaluating peers and that the criterion they used helped themselves prepare their own presentations as a goal setting. Therefore, it can be said that within this group of raters they as a whole functioned with each other as a rater as well as a learner.

Judging from the informal interviews with the students after the presentation class, they said that they learned many things from this peer evaluation in the presentation class as follows:

1) the importance of facial expressions to convey the speaker's sincerity
2) the difficulty of keeping eye-contact by avoiding reading the manuscript
3) the necessity of keeping smile to make the audience relaxed
4) the effectiveness of body language or non-verbal expression such as gestures and posture.

Overall, they realized the importance and necessity of smile and body language as one of the effective methods of giving better presentations.

Table 11. **Items Measurement Report**

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu items
210	48	4.4	4.31	.32	.20	.9	0	.8	0	1 sincerity
208	48	4.3	4.27	.25	.20	.7	-1	.8	-1	2 oral fluency
216	48	4.5	4.44	.55	.19	.5	-3	.5	-2	3 pronunciation
211	48	4.4	4.33	.36	.20	1.1	0	1.0	0	4 eye contact
193	48	4.0	3.96	-.36	.21	1.3	1	1.3	1	5 facial expression
205	48	4.3	4.20	.13	.20	1.1	0	1.1	0	6 grammar
196	48	4.1	4.02	-.23	.20	.8	-1	.8	0	7 originality
204	48	4.3	4.18	.09	.20	1.1	0	1.0	0	8 content
202	48	4.2	4.14	.01	.20	.9	0	.8	0	9 wtitten fluency
189	48	3.9	3.89	-.53	.21	1.1	0	1.0	0	10 evidence
144	47	3.1	3.09	-.61	.23	1.4	2	1.4	2	11 holistic evaluation
198.0	47.9	4.1	4.08	.00	.20	1.0	-.2	1.0	-.3	Mean (Count: 11)
18.7	.3	.4	.35	.36	.0	.2	1.4	.2	1.3	S.D.

RMSE (Model) .20 Adj S.D. .30 Separation 1.49 Reliability. 69
Fixed (all same) chi-square: 34.1 d.f.: 10 significance: .00
Random (normal) chi-square: 9.9 d.f.: 9 significance: .36

6. Conclusion

The following conclusions can be drawn. Firstly, the Rasch based analysis provides us with 1) the relationship among three facets (raters, students, items), 2) the rater severity and fit statistic, 3) the students' ability and fit statistic 4) the item difficulty and fit statistic, and 5) the functioning of rating categories. With all or part of these pieces of information, the facets of the test can be thoroughly investigated individually, which was not be possible in the traditional test analysis.

Secondly, the test can be improved by examining the fit statistics (misfit items) statistically. Also, the test can be improved by having a discussion with raters or students when they are misfitting ones.

Lastly, it can be said that the teacher and peer assessment in presentation class functions for three reasons. One is that students recognize the target level of the presentation for the rating criterion which the teacher set up. The more ratings they do, the more accurate their evaluations become.

Another reason is that by taking turns (presenters and evaluators) they must always concentrate on the in-class activities. In other words, two jobs (rating and presenting) always had students participate in the classroom activities.

Still another reason is that students can learn how to present from the viewpoint of testing or assessment when rating happens in the classroom situation. The better the test becomes organized, the more impact the test has upon the students.

In summary, 1) Peer assessment can be successful in motivating students to improve their presentations. 2) Students can function as reasonably reliable raters of their peers. 3) The Rasch model can be useful in yielding a considerable amount of significant data on several factors involved in presentation assessment, including student academic adequacy, item difficulty, rater severity, and the ability of students to adjust socially.

References

Allison, D. (1999). Language Testing and Evaluation. Singapore: Singapore University Press.

Bachman, L.F, and Palmer, A. S. (1996). Language Testing in Practice. Oxford: Oxford University Press.

Tatum, D. S. (1997a). The Results of a Meaningful Measurement Analysis. Chicago: National Communication Association.

Tatum, D.S. (1997b). Meaningful Measurement: Competency Map — An Item Bank for Speech Evaluation. Chicago, Illinois. Meaningful Measurement.

Tatum, D. S (1999). Persuasive Communication. Chicago; The University of Chicago, Center for Continuing Studies

The Hiyoshi Review Publishing Committee. (2000). Language, Culture and Communication. Special Issue on Foreign Language Education Research. No.24. Keio University.

Wright, B. D, and Stone, M. H. (1979). Best Test Design: Rasch Measurement. Chicago: MESA Press.

Linacre, John M. (1989,1994). Many-Facet Rasch Measurement. Chicago: MESA Press.

Oscarson, Mats (1997). Self-assessment of foreign and second language proficiency. In C. Clapham and D. Corson (Eds). Encyclopedia of Language and Education, Volume7: Language testing and assessment, 175-187.