

## **CLASS ASSESSMENT: CAN STUDENTS BE RELIED ON?**

**By**

Nuraihan Mat Daud and Nor Lide Abu Kasim  
International Islamic University Malaysia

### *Abstract*

One of the educational objectives is to produce students who are critical of their own performance. This can be achieved if they are allowed a more active role in the evaluation process. This study examines the practicality of having students as one of the assessors. The study was conducted at the International Islamic University Malaysia in classes where problem-based learning approach was adopted. Marks given by three categories of assessors namely teacher, self and peer were compared to see whether there were any significant differences among them. Both quantitative and qualitative techniques were used in collecting data. All assessors used the same assessment profile, and diary entry and interview techniques were also used to get insight into the issue. The quantitative data was analyzed using many-faceted Rasch measurement model. The study shows that there were significant differences in the rating given by the three assessors. Different level of severity/leniency was also observed when different tasks were analyzed. The study also looks at the level of difficulties of the criteria used in the assessment. The findings indicate that there were significant differences in the difficulty level of the criteria used. The diary and interviews conducted at the end of the semester revealed that students' had reservation about having to assess themselves. However, the quantitative data shows that the rating given by the students were consistent and not affected by their apprehensiveness.

### **INTRODUCTION**

The quality of learning may be affected by the approach used in teaching the subject. One way of ensuring that it has a life-long effect is to concentrate on skills other than content. Students who can think critically and creatively and possess key competencies such as communication, IT and the ability to work with others and to solve problems are the preferred workers (Mayer Committee, 1992; Mayer Committee, 1992; The Conference Board of Canada, 2003; The Department of Education and Skills, UK, 2003). They may be better able to apply what they have learnt and may cope better with new situations in the workplace.

Many actions can be taken at the point of delivery to ensure that they acquire these skills. One of them is the adoption of peer- and self-assessment. It does not only allow students to be critical of their own learning but also help to develop the needed key competencies if conducted in the right environment. This study investigates the potential for improving quality in teaching and learning at the point of delivery. It focuses on peer- and self-assessment within the context of problem-based learning.

### **STUDIES ON SELF-ASSESSMENT**

Falchikov and Boud (1989) conducted a meta-analysis on student self-assessment in higher education, and found that three factors are closely related to closer correspondence between self- and teacher assessment. They concluded that studies with better design have closer correspondence between teacher and student-assessment, students doing advanced courses (at least three years enrollment) are better assessors than those in introductory courses, and assessments done on studies within the area of science more accurate than other areas. The analysis also shows that the more experienced students tended to underestimate their performance. Another aspect that was found to affect assessment was students' level of anxiety. MacIntyre, Noels and Clément (1997: 265) report that "...anxious students tended to underestimate their competence relative to less anxious students, who tended to overestimate their competence".

Other than factors affecting students' assessment there were also those who looked at the effect of employing self-assessment. Stallings and Tascione (1996), for example, concentrated on Mathematics students in their study. They claimed that self-assessment helped them to be more independent in learning mathematics. They suggest that it can improve students' confidence in doing their subject. Zoller et al (1999) did their study on chemistry students. They state that there is potential for self-assessment. Their study shows that there were differences in the grading made by the professors and the students. The gap was fairly small and statistically significant in lower-order cognitive skills (LOCS)-oriented chemistry examinations but relatively large where higher order cognitive skills (HOCS) was concerned. They state that students' familiarity with a problem and conceptual understanding of it help to make them perform better. These factors also influenced the assessments made. Their self-assessment tended to match the assessment of their professors with greater familiarity and understanding of the subject.

### **STATEMENT OF THE PROBLEM**

A common scenario in the context of this study is that students are employed right after they graduated, and many of them joined the teaching line. As the number of English majors is rather small many are employed even without any training in teaching. Hence, there is a need to include some aspects of teaching at the point of delivery to prepare the students for the workforce. Peer and self-assessments were introduced for this purpose. The experience may help to make them better assessors. Such an experience may also help them be better workers in whatever profession that they chose. Being critical of their own and their colleagues' performance may help make them better personnel.

Apart from the above, peer- and self-assessment were seen as a way of getting students more involved in the learning process. Based on years of teaching experience students were observed to be passive in the class when the traditional mode of teaching was used. Malaysian students are known to be passive (Martin, 1998; Galea, 1999; Nora, 1997). It is very rare that they challenge the views of their teachers. The examinations were the best clue of the students' understanding of the subject. Getting the students to present on the given topic, and assess their own and their peers' performance is a way of encouraging them to take control of their own learning.

## OBJECTIVES OF THE STUDY

The objective of this study is to see whether students' assessment is a reliable mean of assessing their own performance. The study seeks to answer the following questions:

- i. Can language students perform self-assessment and peer-assessment?
- ii. Is their assessment compatible with that of their lecturer and peers?
- iii. Are the students confident in doing the self- and peer-assessment?

## METHOD

### (i) *Subjects of Study*

The study was conducted at the International Islamic University Malaysia. A total of four classes were observed. Two of them were observed in Semester II 2003/4 session, and the other two in Semester I 2004/5 session. Table 1 gives the no. of student in each class:

Table 1: Number of students

Cohort	Class	No. of students	Total no. of students
I (Sem. II 2003/4)	A	20	31
	B	11	
II (Sem. I 2004/5)	C	23	36
	D	13	

The sample of the study was made up of mostly final year students who were majoring in English Language and Literature. The approach was tried in *Computer Applications in Language Studies* classes where students seemed to have difficulty in understanding the content, particularly those that they perceived as technical. Their lack of familiarity with the subject and the technology may have made the course seemed difficult to the students. The adoption of self-assessment was hoped to improve their understanding of the subject as well as develop life-long learning skills. Class periods lasted 90 minutes on alternating days two days a week. The classes were held in a computer lab where each student had access to one computer (with internet connection).

### (ii) *Scoring materials and procedures*

Teacher assessment is taken as the yardstick in this study. The degree of success is measured by the degree of agreement between teacher and student ratings (Falchikov and Boud, 1989). To ensure a greater degree of success a checklist was given to the students. They were taught the points to look for and what each grade means. The assessors for the present study were the teacher, the class and the students themselves. The same rating scales were used by the teacher and the students.

The tasks were scored using a modified tool developed by Alvarstein (2001). Language was included in the modified version since the students were English majors. Each task was scored on six criteria by the three types of raters. Table 2 below is a sample of the assessment sheet used in evaluating the presentations:

Table 2: Checklist used in assessing students' presentation

No.	Criteria	Percentage	Percentage Gained
1.	Communicated the message properly	15%	
2.	Presented a competent discussion about the knowledge achieved	15%	
3.	Presented a satisfactory written abstract	15%	
4.	Defended the solution to the problem at hand	15%	
5.	Explained related areas not covered in the problem statement based on the learning objectives	15%	
6.	Identified references and sources	5%	
7.	Involved the audience	5%	
8.	Language	15%	

Students were divided into groups consisting of between three to four students each at the beginning of the semester. Peer and self-assessment exercises were completed three times during the semester. Problems were given for students to think rigorously on the topics. The problems gave them the opportunities to study issues which were pertinent to their discipline. Ample time was given to the groups to conduct computer and library searches. The students were asked to present the problems and find possible solutions to them. Each topic was followed by a PBL problem module, "with the expectations that students will synthesize the associated content and concepts" (Alexander et al, 2002). Forty percent of the evaluation consisted of PBL and group performance.

An explanation on each topic was given to the students before the problems were presented to them. The evaluation form served as a guide in preparing for their presentation. Students were encouraged to consult the lecturer whenever they found difficulties in looking for the materials. The internet connection in the class facilitated the adoption of this approach in the class. No two groups were allowed to present on the same case.

Before they present, a form was distributed to each individual student to assess the group performance. The presenters were also asked to assess their own performance. The teacher also assessed the same presentation using the same checklist. This is to facilitate comparison among the three assessors.

#### *(iii) Diary Entries*

Students were asked to write what they felt about the class after every lesson in their personal diaries. This is to find the answer to research question number three, that is whether they have the confidence to do the assessment. This, however, was not made explicit to the class as the teacher wanted to give them the freedom to express themselves.

#### *(iv) Interviews*

Students were interviewed after the quantitative results were analysed to get their views on the assessment procedure. It was conducted on a few selected students after the examination was over to reduce the possibility of them holding vital information

for fear that their marks will be reduced if they were to give negative opinions on the subject.

*(v) Data*

The data for this study were ratings on the various presentation tasks, the diary entries and the interviews. In the first cohort, however, the third presentation could not be analysed as many were absent on the day when they were not presenting. Rasch multifaceted analysis was used to see whether the assessments made by the various assessors were statistically significant. Its use for the investigation of rating behaviour in the field of language assessment is relatively recent. Although criticisms against its suitability have been raised particularly with regards to the issue of unidimensionality (e.g., Nunan, 1988 and Hamp-Lyons, 1989), its capacity “to tease out the complexities that the contextual richness of performance assessment introduces” (McNamara, 1996, p. ) makes it an appropriate statistical tool for this study. Hence, in the present study, the Rasch many-facet model and its associated computer application, FACETS (Linacre, 1991-2002), were utilized to investigate the way in which the different groups of raters rate a group of students’ performance on two tasks.

## **RESULTS**

### **FIRST COHORT**

#### **FACETS Summary**

Figure 1 shows the logit scale with 4 types of information: (1) distribution of students in terms of their ability, (2) the severity level of each group of assessors, (3) difficulty of the criteria used in the assessment and (4) the most probable rating categories used in the assessment. The first column is the logit scale followed by student ability distribution, assessor group/ type severity distribution, criteria difficulty information and lastly, the most probable categories used in the rating process.

Measr	+examinee	-Raters	-Items	S.1
+ 1 +				+(10) +
	***			---
	***			
	*****			8
	*****			---
	*****	peer-assessment		7
	**		defend explain	
* 0 *		* self-assessment	* discuss language	* --- *
	*		abstract message	6
		teacher-assessment		---
				5
				---
				4
				---
				3
				---
+ -1 +				+(1) +
Measr	* = 1	-Raters	-Items	S.1

Figure 1: All Facet Vertical “Rulers”: Student ability, Assessor Severity, Criteria Difficulty and Rating Categories Used for Task 1

A) Student Ability

Students are ordered with the most able at the top and the least able at the bottom of the scale. In this analysis, the most able student has a logit of +0.70 and the least able has a logit of -.13. It is evident that there is not much variation of ability among students as the spread of student ability is only about one logit. The person separation index of 2.69 and the chi-square value of 366.1 with 35 d.f. at  $p < .00$  however indicates that students consistently differ from one another in overall ability (See Table 2). For this task the least able student is student no.14 (-.13 logit) whereas the most able students are students no.19 and 15 (both at +.70 logit). The mean measure for this sample of students is 0.39 indicating that they are of a quite high ability. Overall, the standard error of measurement for all students is very small ranging from .04 to .07. Student no. 14 however has quite a large standard error (.20) as the number of observations for this student is small (6 observations). Table 3 also shows that two students (no. 32 and 3) are misfitting; their infit and outfit mean square values outside the recommended range of 0.7 to 1.3 (Linacre, 1991-2004). This means that the ratings given for these students were not consistent with what is expected by the model.

Table 3: Examinee Measurement Report for Task 1

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu examinee
39	6	6.5	5.69	-.13	.20	.35	-1.3	.35	-1.3	.93	14 14
791	131	6.0	6.77	.12	.04	1.32	2.6	1.31	2.4	.62	32 32
791	131	6.0	6.77	.12	.04	1.39	3.1	1.39	3.0	.53	3 3
772	124	6.2	6.92	.16	.04	.91	-.7	.92	-.6	.99	1 1
772	124	6.2	6.95	.17	.04	.94	-.4	.95	-.3	.95	23 23
781	124	6.3	6.98	.17	.04	.93	-.5	.94	-.4	.95	12 12
844	131	6.4	7.11	.21	.04	.74	-2.3	.83	-1.4	1.09	35 35
844	131	6.4	7.11	.21	.04	.70	-2.7	.78	-1.9	1.14	34 34
871	132	6.6	7.24	.25	.04	.79	-1.8	.82	-1.5	1.01	11 11
880	132	6.7	7.29	.27	.04	.78	-1.9	.80	-1.6	1.02	8 8
841	126	6.7	7.34	.28	.05	.83	-1.3	.87	-1.0	1.00	13 13
500	72	6.9	7.41	.31	.06	1.27	1.5	1.29	1.5	.79	5 5
493	72	6.8	7.43	.31	.06	1.20	1.1	1.18	1.0	.88	30 30
500	72	6.9	7.46	.32	.06	1.17	.9	1.16	.9	.91	10 10
913	131	7.0	7.54	.35	.05	1.06	.5	1.04	.3	1.00	18 18
919	131	7.0	7.58	.37	.05	1.09	.7	1.07	.5	.97	20 20
919	131	7.0	7.58	.37	.05	1.11	.8	1.08	.6	.96	31 31
943	132	7.1	7.67	.40	.05	1.16	1.2	1.17	1.2	1.11	16 16
952	132	7.2	7.73	.42	.05	1.09	.6	1.09	.6	1.12	4 4
952	132	7.2	7.73	.42	.05	1.12	.9	1.12	.9	1.08	17 17
582	78	7.5	7.89	.49	.07	1.04	.2	1.03	.2	1.10	9 9
582	78	7.5	7.89	.49	.07	1.08	.5	1.13	.7	1.04	25 25
582	78	7.5	7.89	.49	.07	1.00	.0	1.01	.0	1.14	7 7
582	78	7.5	7.89	.49	.07	1.11	.6	1.08	.5	1.08	29 29
627	84	7.5	7.90	.49	.06	.87	-.7	.86	-.7	1.21	6 6
627	84	7.5	7.90	.49	.06	.83	-.9	.84	-.8	1.26	24 24
627	84	7.5	7.90	.49	.06	.87	-.7	.86	-.8	1.22	33 33
946	126	7.5	7.96	.52	.05	1.25	1.7	1.25	1.6	.73	28 28
946	126	7.5	7.98	.53	.05	1.22	1.5	1.21	1.4	.76	27 27
951	126	7.5	7.99	.53	.05	1.25	1.7	1.24	1.5	.76	22 22
639	83	7.7	8.08	.58	.07	.92	-.4	.92	-.3	1.05	26 26
639	83	7.7	8.08	.58	.07	.90	-.5	.89	-.5	1.07	36 36
639	83	7.7	8.09	.58	.07	.89	-.6	.89	-.5	1.08	2 2
1048	132	7.9	8.30	.69	.06	.86	-.9	.87	-.8	1.17	21 21
1050	132	8.0	8.31	.70	.06	.71	-2.0	.72	-2.0	1.23	15 15
1051	132	8.0	8.32	.70	.06	.75	-1.8	.75	-1.7	1.18	19 19
Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu examinee
762.1	107.9	7.1	7.57	.39	.06	.99	.0	.99	.0		Mean (Count: 36)
206.7	29.8	.6	.54	.18	.03	.21	1.4	.20	1.3		S.D.
RMSE (Model)		.06	Adj S.D.	.17	Separation	2.69	Reliability	.88			
Fixed (all same)		chi-square: 366.1	d.f.: 35	significance: .00							

B) Assessors

The severity level of the different groups of assessors is modeled in the second column. The most severe assessor is placed at the top and the least severe at the bottom (Figure 1). In this analysis, peer assessor is the most severe assessor group (.20 logit) whereas the teacher is the most lenient assessor (-.19 logit). Table 4 gives the measurement report of the three categories of assessors. The range of severity levels is about .40 logit which is about half the range of variability of student ability. The person separation index of 4.04 and the chi-square value of 81.0 with 2 d.f. at  $p < .00$  indicates that the three groups of assessors consistently differ from one another in their judgment of student performance (See Table 3). The infit and outfit mean square statistics will fall in the range of 0.77 to 1.3 indicate that all three categories of raters are consistent in their

scoring/ rating. The standard error of measurement for each assessor category is very small indicating a high precision of the measures.

Table 4: Assessor Measurement Report for Task 1

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Obs %	Agree. Exp %	N Raters
1660	207	8.0	8.08	-.19	.05	.99	.0	.99	.0	.96	17.0	18.2	3 teacher-assessment
1505	197	7.6	7.66	-.01	.04	.92	-.6	.92	-.7	1.09	17.9	18.4	1 self-assessment
24270	3480	7.0	7.04	.20	.01	1.01	.4	1.01	.4	1.00	18.2	17.8	2 peer-assessment
9145.0	1294.7	7.5	7.59	.00	.03	.97	-.1	.97	-.1				Mean (Count: 3)
10695.2	1545.3	.4	.43	.16	.02	.04	.4	.04	.5				S.D.

RMSE (Model) .04 Adj S.D. .15 Separation 4.04 Reliability .94  
 Fixed (all same) chi-square: 81.0 d.f.: 2 significance: .00  
 Rater agreement opportunities: 35640 Exact agreements: 6466 = 18.1% Expected: 6363.5 = 17.9%

C) Assessment Criteria

With respect to the criteria used in the rating of student performance, *explain* is the most severely scored criterion (.15 logit) followed by *defend* (.08 logit). The least severely scored criterion is *message* (-.14 logit) followed by *abstract* (-.07 logit). What this means is that *explain* is the most difficult criterion for students to get high scores whereas *abstract* is the easiest criterion for students to get high scores. The Infit mean-square and Outfit mean-square statistics indicate that all the criteria were consistently rated by the assessors as they all fall within the recommended range of 0.7 and 1.3 logits. The chi-square value of 120.8 with 5 d.f. at  $p < .00$  indicates that the criteria differ in terms of difficulty (See Table 5).

Table 5: Criteria Measurement Report for Task 1

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N Items
4886	650	7.5	7.98	-.14	.02	.99	.0	.99	-.2	1.01	1 message
4734	648	7.3	7.81	-.07	.02	1.25	3.8	1.22	3.3	.87	3 abstract
4667	650	7.2	7.72	-.03	.02	.94	-1.1	.91	-1.5	1.07	2 discuss
4535	644	7.0	7.61	.01	.02	1.19	3.0	1.22	3.4	.93	6 language
4381	645	6.8	7.40	.08	.02	.80	-3.7	.79	-3.9	1.12	4 defend
4232	647	6.5	7.20	.15	.02	.91	-1.6	.91	-1.6	1.00	5 explain
4572.5	647.3	7.1	7.62	.00	.02	1.01	.1	1.01	-.1		Mean (Count: 6)
218.8	2.3	.3	.26	.10	.00	.16	2.7	.16	2.7		S.D.

RMSE (Model) .02 Adj S.D. .09 Separation 4.34 Reliability .95  
 Fixed (all same) chi-square: 120.8 d.f.: 5 significance: .00

D) Rating Categories

In terms of use of rating categories, it is found that only three categories have been used quite extensively. They are categories 5 (C+), 8(B+) and 9 (A-). Figure 2 shows the category probability plot with the categories highlighted. This suggests that assessors identify three main levels of performance.

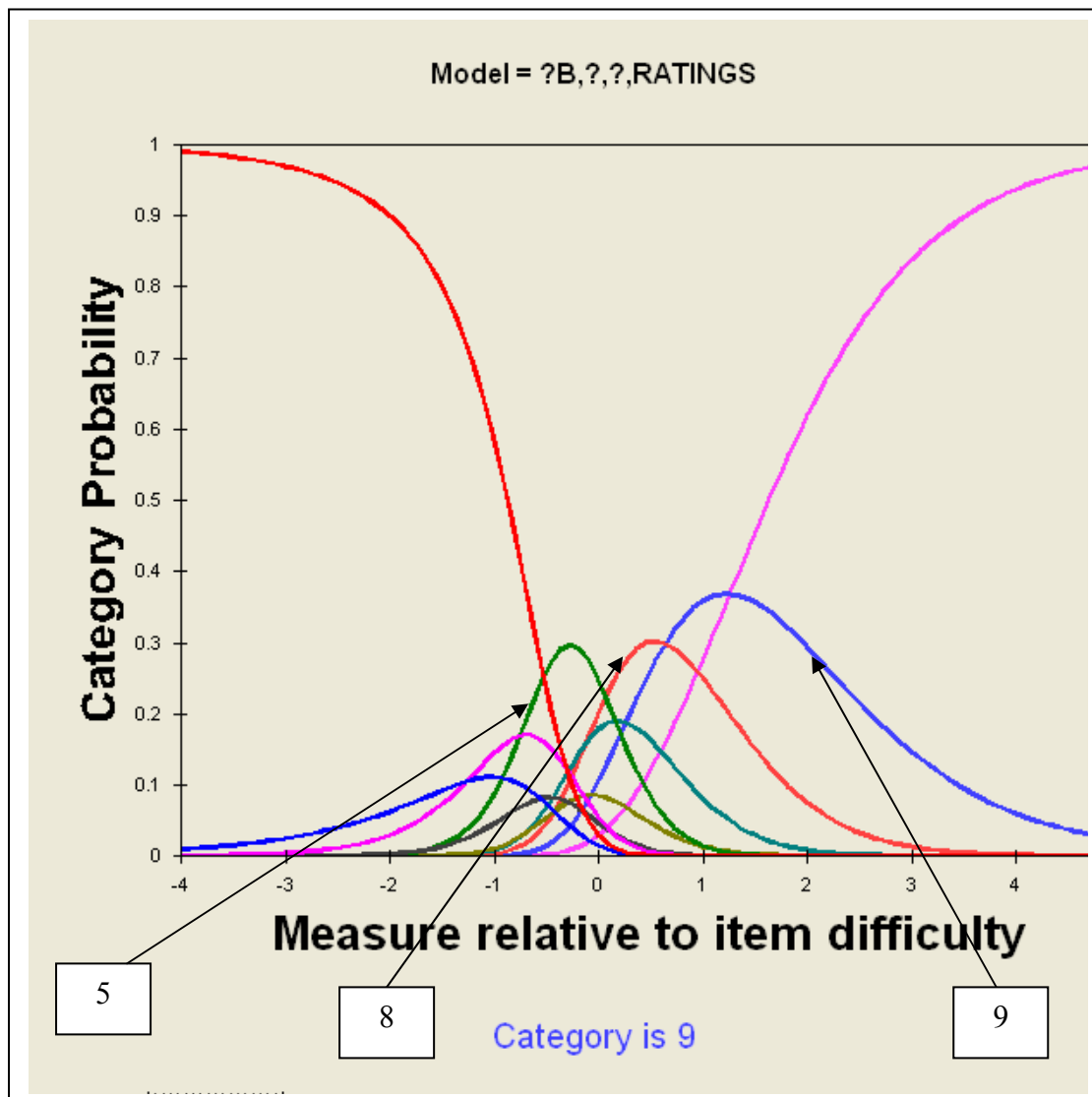


Figure 2: Category Probability Plot for Task 1

The same analysis was conducted for Task 2 with somewhat similar results. However in this analysis there were 33 students instead of 36. Three of the students had very few ratings for the second task and were therefore dropped from the analysis.

#### A) Student Ability

In this analysis, the most able student has a logit of +0.60 and the least able has a logit of -.04 (Figure 3). For this task there is less variability in terms of student ability. However the person separation index of 2.19 and the chi-square value of 237.7 with 32 d.f. at  $p < .00$  indicates that students consistently differ from one another in overall ability (See Table 5). For this task the least able student is student no.35 (-.04 logit) whereas the most able students are students no.36 and 26 (both at +.60 logit). Overall, the standard error of measurement for all students is still very small ranging from .05 to .09. None of the students have a large standard error. Table 6 also shows that three students (nos. 33, 24

and 6) are misfitting; their infit and outfit mean square values outside the recommended range of 0.7 to 1.3 (Linacre, 1991-2004). This means that the ratings given for these students were not consistent with the expectations of the Rasch model.

Measr	+examinee	-Raters	-Items	s.1
+ 1 +				+(10) +
	**			---
	**.			8
	**			---
	*****.			7
	*.			---
* 0 * *		peer-assessment	defend explain	* --- *
		teacher-assessment	discuss language message	6
		self-assessment	abstract	---
				5
				---
				4
				---
				3
				---
+ -1 +				+(1) +
Measr	* = 2	-Raters	-Items	s.1

Figure 3: All Facet Vertical “Rulers”: Student ability, Assessor Severity, Criteria Difficulty and Rating Categories Used for Task 2

Table 6: Examinee Measurement Report for Task 2

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu examinee
653	110	5.9	6.27	-.04	.05	.98	-.1	1.00	.0	.70	35 35
653	110	5.9	6.27	-.04	.05	.93	-.5	.97	-.2	.76	34 34
367	54	6.8	7.01	.17	.07	1.04	.2	1.03	.2	1.39	10 10
414	60	6.9	7.05	.18	.07	1.23	1.2	1.30	1.5	1.09	5 5
366	54	6.8	7.06	.19	.07	1.11	.6	1.10	.5	1.26	30 30
549	78	7.0	7.26	.25	.06	1.36	2.0	1.31	1.7	.96	33 33
549	78	7.0	7.26	.25	.06	1.35	1.9	1.31	1.7	.98	24 24
549	78	7.0	7.26	.25	.06	1.38	2.1	1.33	1.8	.94	6 6
722	102	7.1	7.31	.27	.06	.74	-1.9	.74	-1.9	1.17	18 18
718	101	7.1	7.34	.28	.06	.82	-1.2	.89	-.7	1.02	13 13
770	108	7.1	7.36	.29	.06	.76	-1.8	.75	-1.8	1.15	20 20
770	108	7.1	7.36	.29	.06	.77	-1.7	.76	-1.7	1.14	31 31
385	54	7.1	7.38	.30	.08	1.17	.8	1.18	.9	.80	29 29
346	48	7.2	7.38	.30	.09	1.17	.8	1.29	1.3	.64	25 25
768	107	7.2	7.40	.30	.06	.77	-1.7	.78	-1.5	1.10	11 11
778	108	7.2	7.42	.31	.06	.90	-.6	.89	-.7	.98	12 12
778	108	7.2	7.42	.31	.06	.87	-.9	.86	-.9	1.02	1 1
768	107	7.2	7.42	.32	.06	.78	-1.6	.80	-1.4	1.09	8 8
346	48	7.2	7.45	.32	.08	1.02	.1	1.04	.2	.80	9 9
778	108	7.2	7.45	.33	.06	.89	-.7	.89	-.7	.99	23 23
397	54	7.4	7.58	.38	.08	1.05	.3	1.06	.3	.85	7 7
798	108	7.4	7.59	.38	.06	1.08	.5	1.05	.3	.93	22 22
798	108	7.4	7.59	.38	.06	1.08	.5	1.07	.4	.93	28 28
798	108	7.4	7.62	.39	.06	1.06	.4	1.04	.3	.95	27 27
871	113	7.7	7.88	.51	.06	1.21	1.3	1.18	1.1	.81	19 19
871	113	7.7	7.88	.51	.06	1.19	1.2	1.16	1.0	.83	21 21
883	114	7.7	7.91	.53	.06	.94	-.3	1.00	.0	1.11	16 16
883	114	7.7	7.91	.53	.06	.92	-.4	.94	-.3	1.13	4 4
878	113	7.8	7.93	.54	.06	1.20	1.2	1.16	1.0	.88	15 15
846	108	7.8	7.99	.57	.07	.96	-.2	1.01	.1	1.10	17 17
568	72	7.9	8.03	.59	.08	1.14	.7	1.09	.4	1.17	2 2
616	78	7.9	8.04	.60	.08	1.14	.7	1.09	.5	1.14	36 36
616	78	7.9	8.04	.60	.08	1.10	.5	1.04	.2	1.18	26 26
Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Nu examinee
662.1	91.2	7.2	7.46	.34	.07	1.03	.1	1.03	.1		Mean (Count: 33)
178.2	23.3	.5	.42	.16	.01	.18	1.1	.17	1.1		S.D.
RMSE (Model)		.07	Adj S.D.	.14	Separation	2.19	Reliability	.83			
Fixed (all same)		chi-square: 237.7		d.f.: 32	significance: .00						

B) Assessors

For this task peer assessor is still the most severe assessor group (.09 logit) whereas self-assessment is most lenient (-.13 logit). Table 6 gives the measurement report of the three categories of assessors. The person separation index of 2.02 and the chi-square value of 16.3 with 2 d.f. at  $p < .00$  indicates that the three groups of assessors consistently differ from one another in their judgment of student performance (See Table 7). However, the range of severity levels between the three assessor groups is smaller than that for Task 1 (0.21 logit). This suggests that the three assessor groups are more in agreement in the ratings that they gave for Task 2 than Task 1.

Table 7: Assessor Measurement Report for Task 2

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Obs %	Agree. Exp %	N Raters
1254	162	7.7	7.79	-.13	.05	1.36	2.5	1.29	2.0	.84	20.0	20.6	1 self-assessment
1410	192	7.3	7.40	.04	.04	.86	-1.3	.81	-1.7	1.00	16.2	19.9	3 teacher-assessment
19186	2656	7.2	7.26	.09	.01	1.00	.0	1.01	.3	1.01	19.4	19.7	2 peer-assessment
7283.3	1003.3	7.4	7.48	.00	.04	1.07	.4	1.04	.2				Mean (Count: 3)
8416.7	1168.7	.2	.23	.09	.02	.21	1.6	.20	1.6				S.D.

RMSE (Model) .04 Adj S.D. .08 Separation 2.05 Reliability .81  
 Fixed (all same) chi-square: 16.3 d.f.: 2 significance: .00  
 Rater agreement opportunities: 22882 Exact agreements: 4399 = 19.2% Expected: 4513.9 = 19.7%

C) Assessment Criteria

A similar pattern of criteria difficulty can be seen with respect to the criteria used in the rating of student performance with the exception of the most leniently scored or the least difficult criterion. For this task the least difficult criterion is *abstract* (-.16 logit). Similar to the first task, there is no misfitting criteria for this task as the infit mean-square and outfit mean-square statistics fall within the recommended range of 0.7 and 1.3 logits.

Table 7: Criteria Measurement Report for Task 2

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	N Items
3816	496	7.7	7.86	-.16	.03	1.20	2.5	1.17	2.1	.78	3 abstract
3706	504	7.4	7.57	-.03	.03	.83	-2.6	.86	-2.2	1.12	1 message
3696	504	7.3	7.56	-.02	.03	.75	-4.0	.75	-3.9	1.15	2 discuss
3659	504	7.3	7.49	.00	.03	1.20	2.7	1.19	2.6	.98	6 language
3519	501	7.0	7.28	.08	.03	.93	-1.1	.96	-.6	1.09	4 defend
3454	501	6.9	7.15	.13	.02	1.15	2.2	1.16	2.3	.88	5 explain
3641.7	501.7	7.3	7.48	.00	.03	1.01	-.1	1.01	.1		Mean (Count: 6)
121.1	2.9	.3	.23	.09	.00	.18	2.7	.17	2.5		S.D.

RMSE (Model) .03 Adj S.D. .09 Separation 3.24 Reliability .91  
 Fixed (all same) chi-square: 66.1 d.f.: 5 significance: .00

In terms of variability, the range of severity for the assessment criteria is the same (.29 logit) as Task 1 (Table 8) below:

Table 8: Level of Difficulty for each criteria

No.	TASK 1	TASK 2
1.	Message	Abstract
2.	Abstract	Message
3.	Discuss	Discuss
4.	Language	Language
5.	Defend	Defend
6.	Explain	Explain

*D) Rating Categories*

In terms of use of rating categories, it is found that the same three categories (5 (C+), 8(B+) and 9 (A-)) have been used more extensively than the rest extensively. Figure 4 shows the category probability plot with the categories highlighted. This suggests that assessors identify three main levels of performance. However, category 8 is not as well-defined as in Task 1. This in a way suggests that only three categories of performance are discernible to the assessors.

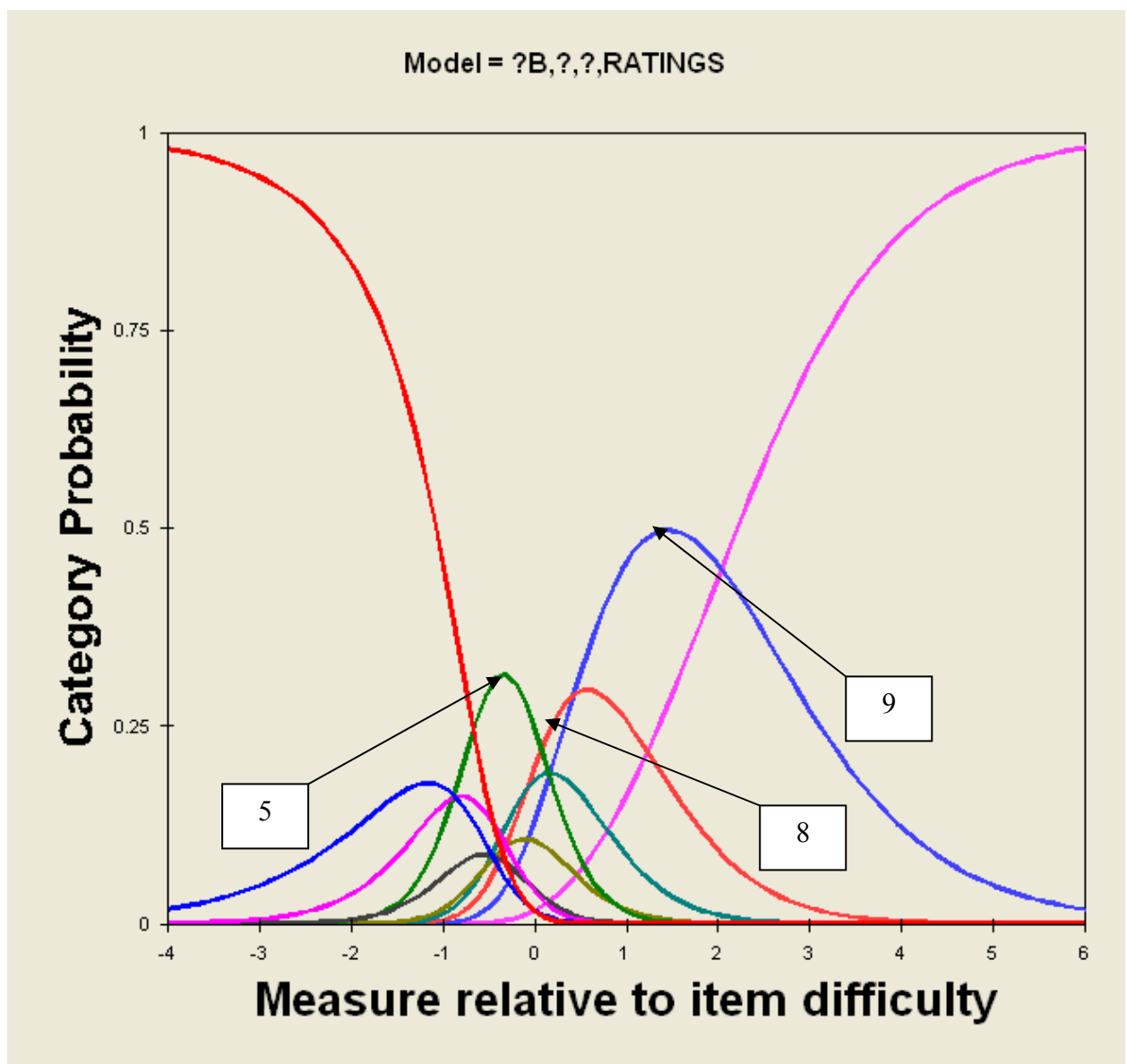


Figure 4: Category Probability Plot for Task 2

**SECOND COHORT:**

To further examine the judging behaviour of the three groups of assessors, a similar ...was ...with another cohort. This cohort comprised.....

The results of the analyses of the judging behaviour of the three assessor groups indicate the following:

*A) Student Ability*

Student ability, though varied, spans about one logit. This shows that the ability range of the students in the second cohort like the first are not very different.

*B) Assessors*

In terms of assessor category, the three groups differed in their judging ( $p \leq 0.00$ ). However the ordering of the three groups with regard to judging severity is somewhat different from the first cohort. For the second cohort:

Table 9: Order of assessors' leniency

TASK 1	TASK 2	TASK 3
Self-assessment (-.17 logit)	Self-assessment (-.10 logit)	Self-assessment (-.29 logit)
Peer-assessment (.02 logit)	Peer-assessment (.02 logit)	Peer-assessment (.02 logit)
Teacher-assessment (.15 logit)	Teacher-assessment (.08 logit)	Teacher-assessment (.27 logit)
Range: .34 logit	Range: .20 logit	Range: .58 logit

- Order of severity remained the same for all three tasks: self-assessment least severe and teacher assessment – most severe.
- Greater disparity in severity is evident in Task 3
- Range of disparity is smallest in Task 2

*C) Assessment Criteria*

Table 10 shows a similar trend with the first cohort in terms of ordering of criteria.

Table 10: Level of Difficulty for each criteria

TASK 1	TASK 2	TASK 3
Message	Abstract	Abstract
Language	Message	Discuss
Abstract	Language	Message
Discuss	Discuss	Language
Defend	Explain	Defend
Explain	Defend	Explain

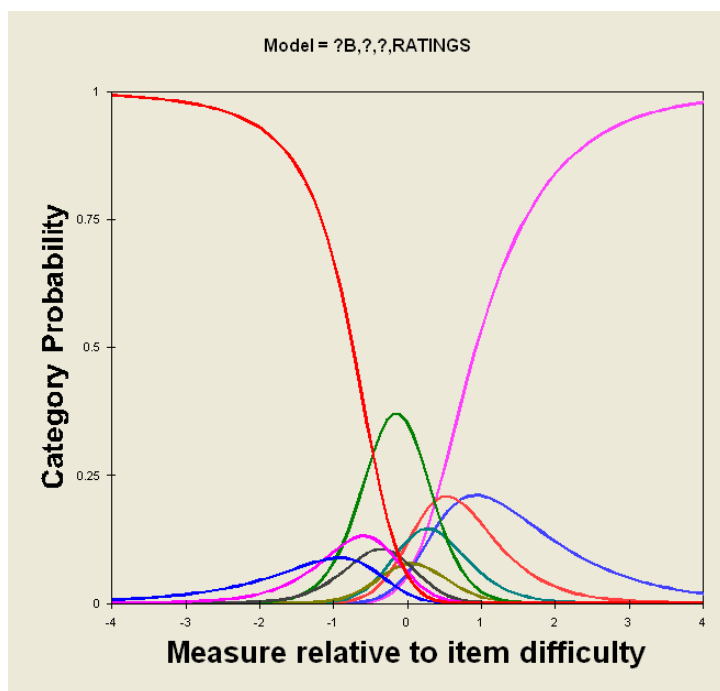
- No misfitting criterion.
- Like in the first cohort, 'Defend' and 'Explain' are the two most difficult criteria whereas 'Message' and 'Abstract' are the easiest.

*D) Rating Categories*

Use of rating categories:

- The same for all three tasks: Category 5 (equivalent to C+) is the most used category

- Different from the first cohort; three categories that were frequently used: Category 5 (equivalent to C+), Category 8 (equivalent to B+), and Category 9 (equivalent to A-).



### DISCUSSION:

Results indicate that:

- Assessors: The three assessor groups differed in terms of judging severity. They however are consistent in their ratings as indicated by the infit and outfit statistics.
- Assessment Criteria: The ordering of the assessment criteria is somewhat similar with 'Defend' and 'Explain' as the most difficult criteria and 'Message' and 'Abstract' are the easiest ones to get high ratings.
- Rating categories: Not all rating categories are frequently used by raters.

### ANALYSIS OF OTHER DATA

The qualitative data revealed what the students felt about the approach. Those who were asked felt that they improved with every presentation. The following excerpts (taken from the students diaries) reflect their feeling:

- (1) Student 1: The advantage of having frequent presentation is that it helps us to bring perfection in our presentation skills. From a shy person, we are mould to be a confident speaker through consistent presentation (sic).
- (2) Student 2: My third presentation, I found that I was not having any difficulties in front of the class and there was no more anxiety. I felt quite confidence actually (sic).

There were students who were troubled by the quality of their presentation. They were not sure how the others took it. Some felt that they let their group members down when they could not explain clearly. There were those who felt frustrated when

their classmates gave them a blank look. This was especially the case with their first presentation. They experienced a high level of anxiety then. But as they gained their confidence and they were more familiar with the subject they might feel that they deserved to get a better grade. Excerpt 3 may represent what majority of the students felt in the first presentation and excerpt 4 may be one of the explanations why they were less severe in their assessments of themselves in the consequent task:

- (1) Student 3: I admit that I was not able to measure my own work....Finally I ended up giving average marks as to be fair to myself.
- (2) Student 4: As we tend to grade ourself as what we think we deserve after all the hard works we put in (sic).

The interviews revealed that students tended to include their effort when evaluating themselves. This may also another reason for their leniency as more tasks are assigned to them.

Basically most of the students were very grateful for being given the opportunity to grade themselves. They said:

- (1) Student 5: This is a good way to tell the teacher about what grade we think we deserved.
- (2) Student 6: Fair grading is also achievable because the role does not fall solely on the lecturer alone.

Where peer-presentation was concerned, one student (student 7) said:

I would evaluate based on the assessment form, which will also determine how much have I understood from their presentation and how successful they are in attracting the audience's attention.

Feelings such as this could be the reason for their consistent marking.

## **DISCUSSION**

The analysis shows that the raters were all consistent in their marking. The difference lies in the leniency of their marking. It was observed that the students were more severe in the first task. As in Falchikov and Boud (1989), MacIntyre, Noels and Clément (1997) and Zoller et al (1999) observations, their lack of familiarity and anxiety may be among the reasons for the severity in assessment. As they became more familiar and less anxious they

## **CONCLUSION**

Based on the given data we may conclude that self-assessment may need to be taken with caution as it is influenced by what the students felt rather than what they presented. Peer assessment may be more reliable as they based their assessment on what was presented using the checklist given. However, with proper training they may be a reliable assessor as the difference in their grading with that of their teacher's was not that big. Given that experienced teachers are not always reliable markers in all situations, then it is perhaps, unreasonable to expect inexperienced student to

demonstrate reliability (Falchikov and Boud, 1989; MacIntyre, Noels and Clément, 1997). Since assessment is a skill, giving them the opportunity to do it may be one of the effective ways of helping them to develop it.

## REFERENCES

- Alexander, J.G., G.S. McDaniel, M.S. Baldwin and B.J. Money (2002) Promoting, applying, and evaluating problem-based learning in the undergraduate nursing curriculum, *Nursing Education Perspectives*, 23 (5), 248-253.
- Alvarstein, Vidar (2001) Problem-based learning approach in teaching lower level logistics and transportation, *International Journal of Physical Distribution and Logistics Management*, 31 (7/8), 557-573.
- Falchikov, Nancy and Boud, David (1989) Student self-assessment in higher education: A meta-analysis, *Review of Educational Research*, Winter, 59(4), 395-429.
- Galea, S.R. (1999) Apprenticeship in thinking and the adult learner, 8<sup>th</sup> *International Conference on Thinking: Thinking for a Change Society*, 5-9 July, Shaw Conference, Edmonton, Alberta, Canada.
- Greenan, H., McIlveen, K. and Humpreys, P. (1997) Involving students in teaching and learning: a necessary evil? , *Quality Assurance in Education*, 5(4), 231-235.
- MacIntyre, Peter D., Noels, Kimberly A. and Clement, Richard (1997) Biases in self-ratings of second language proficiency: The role of language anxiety, *Language Learning*, 47 (2), 265-287.
- Martin, Kenn (1998) Staff reflection on UWA as a foreign investment, *Issues of Teaching and Learning*, 4 (10). Retrieved November 1st 2003 from the World Wide Web at <http://www.catl/uwa.edu.au/NEWSLETTER/issue1098/staff.html>.
- Mayer Committee (1992) *Putting General Education to Work: The key competencies report*, AET/MOVEET, Melbourne.
- Nora Nasir (1997) *ESL Learner Difficulties in the Malaysian Literature Classroom*, Doctoral Dissertation, University of Strathclyde.
- OECD (2000) *Links between Policy and Growth: Cross-country evidence – Working party No. 1 on macroeconomic and structural policy analysis*, Paris.
- Pennington, G. and Cannon, R. (1989) Enhancing the quality of teaching and learning in higher education:

Saunders, E. and Saunders, C. (1995) A empirical approach to the identification of teaching skills in higher education, *Journal of Further and Higher Education*, 19(2), Summer, 98-112.

SCANS (1992) *Learning a Living: A blueprint for high performance*, Department of Labor: Washington.

Stallings, Virginia and Tascione, Carol (1996) Student self-assessment and self-evaluation, *The Mathematics Teacher*, 89(7), 548-554.

The Conference Board of Canada (2003) *Employability Skills 2000+*. Retrieved from <http://www.conferenceboard.ca/education/learning-tools/employability-skills.htm>. Retrieved on Sept. 12<sup>th</sup> 2003

The Department of Education and Skills, UK (2003) *Key Skills*, Retrieved from <http://www.dfes.gov.uk/keyskills/> on Sept. 12, 2003

Zoller, Uri, Fastow Michal, Lubezky, Aviva, and Tsaparlis, Georgios (1999) Students' self-assessment in chemistry examinations requiring higher- and lower-order cognitive skills, *Journal of Chemical Education*, 76 (1), 112-113.