Judging Behaviour And Rater Errors: An Application Of The Many-facet Rasch Model

Noor Lide Abu Kassim
noorlide@iium.edu.my
Institute of Education
International Islamic University Malaysia

Abstract

Of the potential sources of construct irrelevant variance or unwanted variability in performance assessment, those associated with raters have been found to be extensive, difficult to control, and impossible to eliminate. And as rater-related errors are non-trivial and threaten the validity of test results, it is necessary that these errors are accounted for and controlled in some way. This paper explains the different types of rater errors and illustrates how they can be identified using the Many-facet Rasch Model, as implemented by FACETS. It also demonstrates what these errors mean in terms of actual judging or rating behaviour and elucidates how they may affect the accuracy of estimation of performance. Rater errors that are explicated in this paper are those related to rater severity, restriction of range, central tendency, and internal consistency. As assessment and its procedures are central to student learning, matters related to valid and fair testing need to be taken seriously. It is hoped that with greater awareness of how we judge and a better understanding of how rater-related errors are introduced into the assessment process, we can be better raters and better teachers.

Keywords: rater error, judging behaviour, Many-facet Rasch Model, performance assessment, validity.

Introduction

The advent of performance assessment not only brought with it promises of greater validity but also greater risks of unwanted variability (Linacre, 1989; McNamara, 1996; Wilson & Case, 2000). Performance assessment, unlike the traditional fixed-response assessment, has features that are peculiar to its assessment setting – the task choice, the task processing conditions, the raters, the rating scale and the rating procedures that involve subjectivity of human judgment – that make it much more vulnerable to construct irrelevant variance (McNamara, 1996; Upshur & Turner, 1999). Of these potential sources of irrelevant variance or unwanted variability, those associated with raters have been found to be extensive, difficult to control, and impossible to eliminate (Linacre, 1989; Lunz, 1997; McNamara, 1996). And as rater-related errors are non-trivial and threaten the validity of test results, it is necessary that these are modelled, accounted for, and controlled (Linacre, 1989).

ISSN: 1675-8021

Of the different types of rater error, the most widely known and understood is rater severity. Rater severity refers to the tendency for raters to consistently give higher or lower ratings than is justified by the performances (Engelhard, 1994). Differences in rater severity occur when raters do not interpret the rating scale in the same way, or have different standards or expectations. The same performance may be considered to be good, average, or poor by different raters. To identify differences in rater severity, interrater-agreement or reliability is often examined. This is the degree to which raters agree in the ratings that they give. If raters are highly in agreement with one another in their ratings, interrater reliability will be high; if their ratings differ substantially, then interrater reliability will be low.

Central tendency and restriction of range are two other types of rater error. Central tendency happens when middle categories are used predominantly by raters. This judging behaviour often reflects the reluctance to use extreme categories. If ratings are somewhere in the middle categories, there is a good chance that the ratings will not be too far from those given by another rater. Disagreement therefore becomes unlikely as the "implicit rule [is] when in doubt, avoid extreme categories" (Linacre, 1998, p. 631). Cases of central tendency are typically detected by examining the pattern of category usage.

Restriction of range, on the other hand, occurs when ratings are restricted to very few categories. Some raters may overuse the lower end of a scale while others may overuse the upper end. As restriction of range pertains to overuse of certain rating categories, central tendency is, therefore, a special case of restriction of range. These two types of rater error are considered a serious threat to the quality of ratings as they fail to accurately discriminate examinees of different performance levels (Saal, Downey & Lahey, 1980). A very severe or lenient rater may be considered to exhibit this kind or rater error.

Another type of rater error relates to the internal consistency of ratings given by individual raters. Problems of internal consistency can be seen when raters are not consistent or constant in their judgment of similar performances. High ratings should be given to all good performances while low ratings should be given to all poor performances. Sometimes due to fatigue or inattentiveness, raters may award a high rating to a poor performance and a low rating to a good performance. Compared to rater severity, this type of rater error is considered to be more serious as raters are in themselves inconsistent in their judgment (Linacre, 1989).

When raters consistently rate certain sub-groups consistently lower or higher, bias is said to be present. Kondo-Brown (2002) found raters in her study to show significant bias towards certain sub-groups and the percentage is more for high and low ability groups. Some raters consistently award higher scores to low ability groups while others award lower scores to high ability groups. Bias may also happen when raters rate certain criteria more harshly or more leniently. For example, Wigglesworth (1993) found that some raters rate grammar and vocabulary more harshly or leniently than others.

The 'halo effect' is yet another undesirable rater effect that contributes to error in the measurement of performances (Engelhard, 1994; Holzbach, 1978; Saal et al., 1980). A halo effect is said to be present when "a rater fails to distinguish between conceptually distinct and independent aspects of an examinee's [performance]" (Engelhard, 1994, p. 98). This type of rater error can be seen when analytic-type rating scales are used. A typical example of halo effect is when a rater gives the same score for different aspects of a performance.

Dealing with rater-related variability

An important question at this juncture is how do we deal with these rater errors or unwanted variability? Within the Classical Test Theory (CTT), variability as a result of rater errors or effects has largely been controlled through the use of multiple raters. The reliability, i.e., the statistical reproducibility, not the substantive quality, of ratings increases when two or more raters are used in the scoring procedure. With more raters (therefore ratings), the precision in measurement becomes higher as more information is available to estimate a performance. In CTT, it is also demanded that raters agree in their judgment. The more similar the ratings awarded, the higher the level of rater agreement, and the higher the interrater reliability.

Given this requirement, one major source of evidence in determining the reliability of ratings within CTT is the investigation of interrater reliability. However, the notion that interrater reliability – or more accurately, rater agreement – can be a real measure of reliability has been questioned by many (e.g., Engelhard, 1994; Henning, 1997; Linacre, 1989) as it fails to give an "accurate approximation of the true ability score" (Henning, 1997, p. 53). Henning (1997, pp. 53-54) argues,

> …two raters may agree in their score assignments and both be wrong in their judgments simultaneously in the same direction, whether by overestimating or underestimating true ability. If this happens, then we have a situation in which raters agree, but assessment is not accurate or reliable because the ratings fail to provide an accurate approximation of the true ability score. Similarly, it is possible that two raters may disagree by committing counterbalancing errors in opposite directions; that is where one rater overestimates true ability, and the other rater underestimates true ability. In this latter situation, it may happen that the average of the two raters' scores may be an accurate and reliable reflection of true ability, even though the two raters do not agree in their ratings.

Secondly, the expectation that raters should agree in their judgment is difficult to support. No two raters can be perfectly unanimous in their judgment of every performance that they encounter (Engelhard, 1994; Linacre, 1989). The requirement within CTT that raters must agree with one another is also counterproductive. This is explained in Linacre (1998, p.631),

> ...the fact that raters know that agreement is preferable constrains their independence (each rater also considers the other rater when assigning a rating) and leads to deterministic features in the data. ... This induces an artificial security in the reported results. The rating scale is reported to be "highly discriminating", and the ordering of the performances is considered "highly reliable". But all this is illusory. The constraint of forced agreement has mandated it.

Given the limitations of CTT in addressing rater-related variability or error – as well as other measurement issues which are beyond the scope of this paper – there has been a shift towards the use of more robust measurement models (see Engelhard, 1994; Kondo-Brown, 2002; Lumley & McNamara, 1995; McNamara, 1996; Wigglesworth, 1993). One such model that has gained credence is the Many-facet Rasch Model (MFRM), developed by Linacre (1989). MFRM models and adjusts for variability that is introduced in ratings through the use of multiple raters, tasks, and any other facet that constitutes the testing procedure. As the aim of any testing process is to provide fair and accurate estimation of examinee performance, the measure that is given to an examinee must be independent of the particular rater or raters or tasks that are used in the judging process (Linacre, 1989).

MFRM is particularly significant in this respect. It facilitates the "observation and calibration of differences in rater severity making it possible to account for these differences in the interpretation of the assigned rating" (Linacre, Engelhard, Tatum & Myford, 1994, p. 569). In other words, MFRM does not expect raters to rate or judge identically. Instead, it accepts and controls for differences in rater severity (Linacre, 1989). A further advantage of MFRM is that each item can be defined with its own scale, or each judge can be modelled according to the manner he or she uses the rating scale (Linacre, 1989; Linacre et al., 1994). Interactions between facets – which may signal bias – in the testing process can also be modelled and statistically tested. In addition, MFRM is able to detect other rater effects such as restriction of range, halo effect and internal inconsistency through the use of particular fit statistics. The simple general form of MFRM can be expressed as follows (Linacre, 1989):

$$\log \left[\frac{P_{nijk}}{P_{nijk-1}} \right] = B_n - D_i - C_j - F_k$$

Where:
P_{nijk} is the probability of examinee *n* being awarded on item *i* by judge *j* a rating of *k*
P_{nijk-1} is the probability of examinee *n* being awarded on item *i* by judge *j* a rating of *k-1*
B_n is the ability of examinee *n*
D_i is the difficulty of item *i*
C_j is the severity of judge *j*
F_k is the extra difficulty overcome in being observed at the level of category *k*, relative to category *k-1*

The utility of MFRM in handling rater-related variability and errors has been discussed and explicated by a number of authors; however, this has been done largely in the field of language testing and measurement (see Engelhard, 1994; Kondo-Brown, 2002; Lumley &

McNamara, 1995; McNamara, 1996; Wigglesworth, 1993). Given the nature of the discipline and its specialized readership, these papers were rather technical in their treatment of the subject and may not have been easily accessible to those without the relevant technical knowledge. The aims of this paper, therefore, are (i) to demonstrate as simply as possible how the different types of rater errors, namely, rater severity, restriction of range, central tendency and internal consistency can be identified using MFRM as implemented by FACETS; (ii) to illustrate what rater errors mean in terms of actual judging or rating behaviour through the use of simple graphs and plots; and (iii) to elucidate how these errors may affect the accuracy of estimation of student performance.

Methodology

Rater

The 34 raters who participated in this study were English Language instructors at a preparatory centre for a higher education institution in Malaysia. They were predominantly second language speakers and were invited to participate in this study as part of a standardization exercise organized by the Testing and Measurement Unit of the English Language Department. Their academic qualifications ranged from bachelor's to master's degree. Areas of specialization include TESOL, Applied Linguistics, and English Language Literature and the number of teaching experience was no less than a year.

Materials and Method

The materials used for this study were 12 paragraphs written by new-intake students for a placement test conducted by the institution. The length of the paragraphs ranged between 100-120 words. The topic was "My Favourite Game". These paragraphs were selected to represent exemplars of writing at each performance level. In order to eliminate context effects, the 12 paragraphs were randomly ordered for different raters. The paragraphs were holistically scored, and the scoring scale used in the judging of the paragraphs was a 10-point rating scale, with a passing score of '5'. There were no ratings of '1'. The scoring procedure and the rating scale used in this study were similar to those used in the scoring of students' writing in the placement test. As raters were only required to score a small number of writing samples, a complete judging plan was used. This means that all raters were required to rate all the writing samples. There were, however, four missing ratings. Rater 29 did not rate two of the writing samples, and Raters 12 and 24 each did not rate a writing sample. Since MFRM accommodates missing data, no adjustments were needed. A complete judging plan is best as it provides maximum linkage between raters and the writing samples. However, it is not always possible to adopt this type of judging plan, especially when a large number of writing samples have to be scored by a limited number of raters in a very short time. Other judging plans with minimal linkages are typically used in such situations.

Data Analysis

The ratings given by each rater were analyzed using FACETS (Linacre, 2003), a computer application which implements the Many-facet Rasch Model. FACETS was used to model and estimate examinee ability, rater severity, and identify other rater errors. The statistical package, SPSS version 13.0, was used to generate descriptive statistics, the distribution of ratings, and for plotting instructors' ratings and examinee ability estimates derived from the FACETS analysis.

Results

Distribution of raw ratings

Figure 1 shows the distribution of ratings given by the raters for each of the paragraphs. From the boxplots, it is clear that raters differ in the severity of their judgment of the individual paragraphs. The difference in ratings ranges from 3 to 5 points. In terms of median rating, Paragraph 10 has the highest median rating (8 points); Paragraphs 1, 2, 3, 5, 7, and 11 share the lowest median rating (4 points). Figure 1 also indicates the presence of some outlying ratings. These are especially evident for paragraphs 10 and 12. It is interesting that although Paragraph 10 is considered a good paragraph by most raters, there are several raters who gave it very low ratings. This may be due to restriction of range or rater bias. Another important observation has to do with the placement of the ratings in relation to the passing score. As the passing score is a rating of 5, it is clear that only Paragraph 10 has been clearly judged as a clear pass. This indicates that whether a student passes of fails is highly dependent on who is judging him or her.

ISSN: 1675-8021

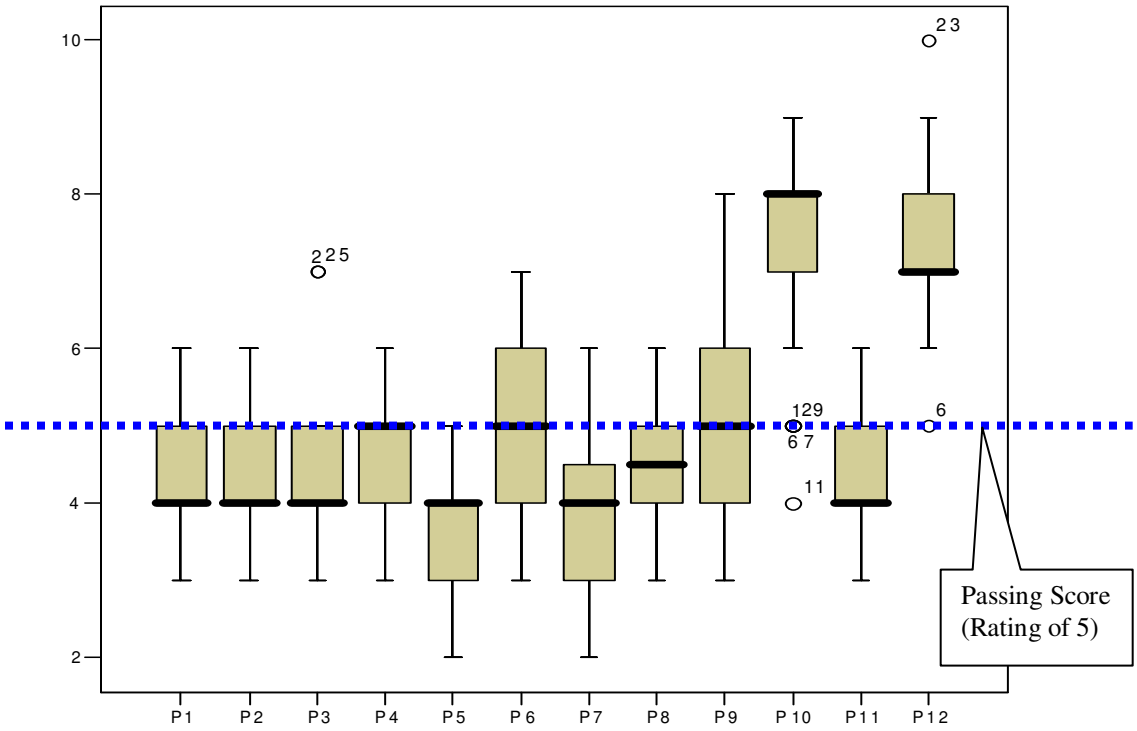


Figure 1: Distribution of ratings for individual paragraphs

FACETS analysis

Figure 2 gives a graphic presentation of ability estimate for each student (i.e. paragraph) and rater severity which is generated by FACETS. The first column on the right is the logit scale, the measurement unit in which student ability and rater severity are measured. The second column gives the distribution of student ability estimates whereas the third column presents the rater distribution. The rater distribution is modelled with a mean of zero, which is the average severity for the raters. Ability measures are adjusted for differences in rater severity and ordered along the logit scale with the most able at the top and the least able at the bottom of the scale. In this analysis, the student with the highest ability estimate (Paragraph 12) has a measure of approximately +3.17 logits, and the student with the least ability estimate (Paragraph 5) has a measure of approximately -2.97 logits. From the distribution, it is evident that there is a considerable amount of variation in ability (a range of about 6 logits). This is desirable as variability in ability is the aim of the measurement process.

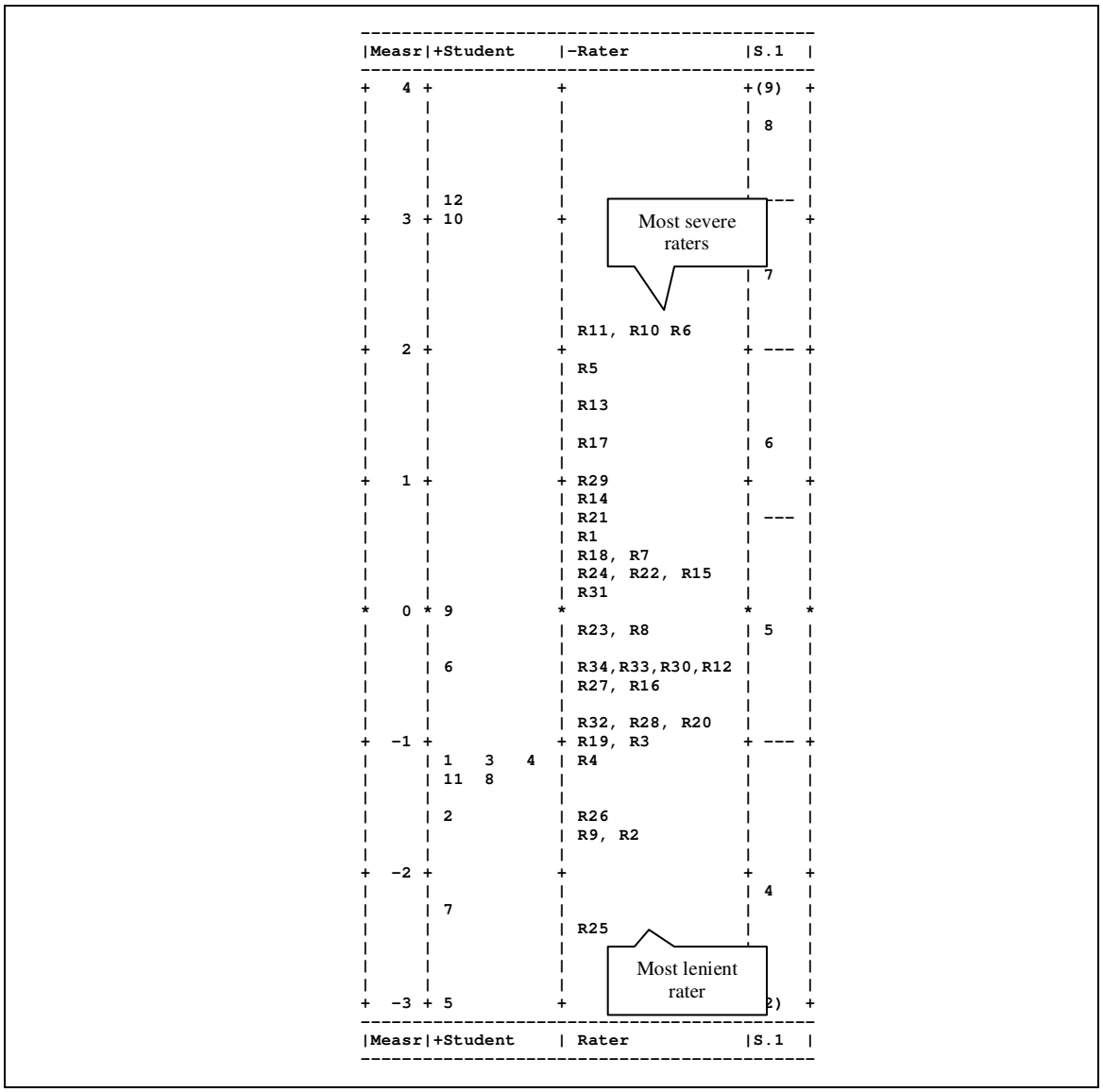


Figure 2: Student ability and rater severity distributions

Rater severity

The severity level of raters is modelled with the most severe rater at the top and the least severe (most lenient) at the bottom of the logit scale. The range of rater distribution is almost as wide as the ability distribution for students/paragraphs. This indicates that these raters differ considerably in their severity level. This also suggests that students' performances would either be grossly underestimated or overestimated if raw ratings (unadjusted for rater severity) are used in the reporting of test results. The right-most column represents the expected average rating. For example, Paragraph 9 (of ability measure 0 logits) has an expected average rating near '5' by raters of average severity, such as 23 and 8. Paragraph 7 (of ability measure -2.3 logits) has a lower average expected rating near '4' by the same raters). The most severe raters are Raters 6, 10, and 11 (+2.09 logits) while

the most lenient rater is Rater 25 (-2.41 logits). Raters 31, 23, and 8 are of average severity. Table 1 gives the raw ratings awarded by the raters for each of the paragraphs. Notice that Rater 11, the most severe rater, gave consistently low ratings for the paragraphs, and Rater 25, who is the most lenient rater, gave higher ratings.

Table 1: Raw ratings by raters

Highest ability ← Lowest ability

Rater	P12	P10	P9	P6	P4	P3	P1	P8	P11	P2	P7	P5
11	6	4	5	4	3	5	4	5	3	3	2	2
10	6	8	3	4	4	4	4	3	3	4	4	3
6	5	5	4	4	3	3	4	4	4	3	3	4
5	7	7	4	3	4	4	4	3	3	3	2	3
13	7	7	4	4	4	3	4	3	4	3	3	3
17	6	7	5	5	3	4	4	3	4	4	3	3
29	.	5	5	.	4	3	4	4	4	5	3	4
14	7	7	4	5	4	5	3	3	4	4	4	3
21	9	8	4	4	3	4	3	5	3	4	4	3
1	6	5	5	5	4	4	5	5	4	4	4	4
18	7	7	3	6	3	4	5	5	4	4	3	5
7	8	5	4	4	5	4	4	3	6	4	4	5
24	9	9	.	4	4	4	4	4	4	4	3	3
22	7	6	6	5	5	5	5	4	4	4	4	3
15	7	8	4	4	4	5	4	6	4	5	4	4
31	6	7	5	5	5	5	5	5	4	5	4	5
23	10	9	8	4	5	4	5	3	5	4	5	3
8	7	8	6	6	4	5	5	4	4	4	3	4
34	7	8	4	5	5	5	4	5	5	5	4	4
33	9	8	6	5	5	5	4	4	4	4	3	3
30	7	9	4	6	5	4	5	5	5	5	4	3
12	8	8	6	4	5	4	.	5	4	4	5	4
27	9	9	6	5	5	4	4	4	4	4	4	4
16	7	7	5	5	5	5	5	4	5	5	4	5
32	7	7	6	5	6	5	5	5	5	5	4	5
28	8	8	6	5	4	4	5	6	4	5	6	3
20	8	7	7	5	5	5	4	4	5	5	5	4
19	8	6	5	6	6	4	5	6	5	5	5	5
3	8	8	6	6	4	5	4	5	6	5	4	4
4	8	8	5	6	5	5	6	4	5	5	5	4
26	8	9	6	6	6	5	6	5	6	5	5	3
9	9	8	5	7	6	5	6	5	5	5	5	5
2	8	9	8	6	5	7	4	6	6	4	4	4
25	9	9	8	6	6	7	6	6	5	4	6	4

Accuracy of estimation of student performance

In any measurement process, accurate estimation of performance is vital for valid measurement. Tables 2 and 3 demonstrate how differences in rater severity have affected the estimation of student performance. Table 2 shows median ratings awarded by raters and the logit measures derived from the MFRM analysis. Table 3, on the other hand, gives the ranking of the students/ paragraphs based on median ratings and MFRM logit measures. Before adjustments were made to differences in rater severity (i.e. based on median rating) Student/Paragraph 10 was ranked first; but after adjusting for rater

severity, Student/Paragraph 12 was ranked first (Table 3). Notice also that median ratings unadjusted for rater severity are unable to discriminate between performances of different ability unlike the MFRM logit measures. Figure 3, further illustrates the effects of rater severity on the accuracy of the estimation of performance.

Table 2: Comparisons between student/paragraph median ratings and MFRM logit measures

	Paragraph											
	1	2	3	4	5	6	7	8	9	10	11	12
Median Rating	4	4	4	5	4	5	4	4.5	5	8	4	7
Rasch Logit Measure	-1.13	-1.64	-1.11	-1.11	-2.97	-.41	-2.35	-1.23	.06	2.98	-1.34	3.17

Table 3: Comparisons between student/paragraph ranking based on median ratings and MFRM logit measures

	Paragraph											
	1	2	3	4	5	6	7	8	9	10	11	12
Ranking (Raw Rating)	5	5	4	3	5	3	5	4	3	1	5	2
Ranking (Rasch Measure)	7	10	5	5	12	4	11	8	3	2	9	1

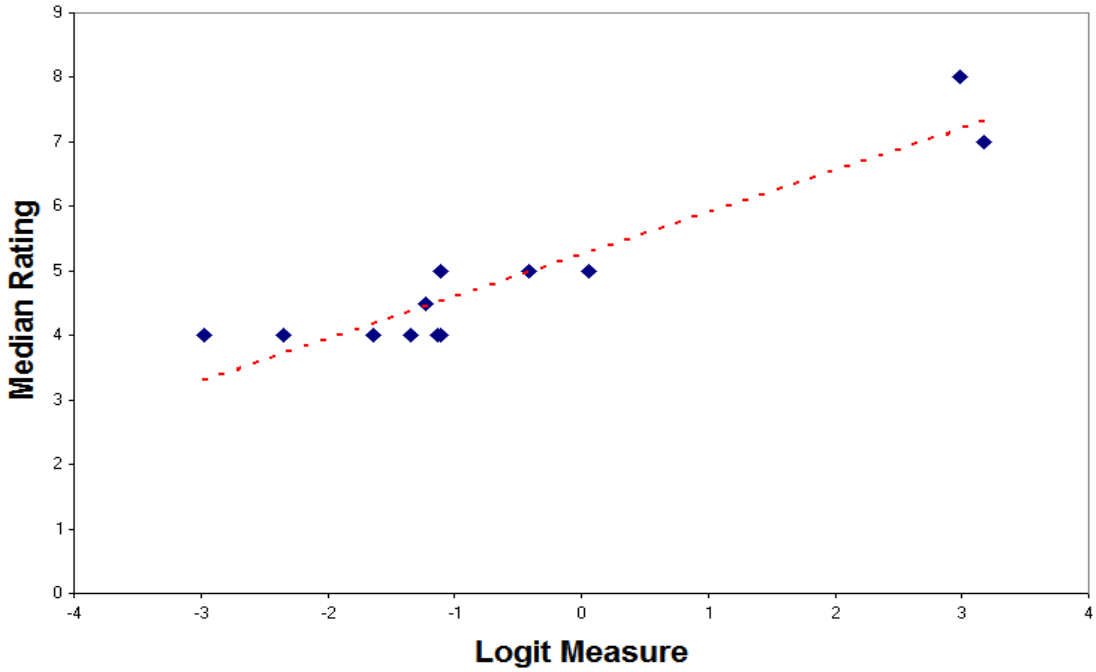


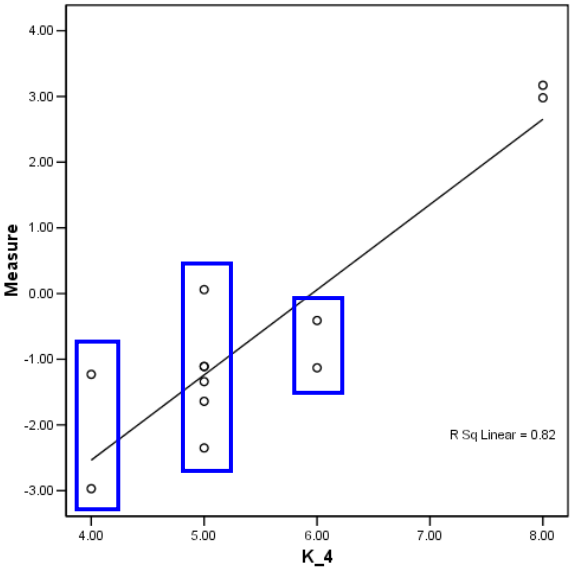
Figure 3: Scatterplot of median ratings and MFRM logit measures

Rater fit statistics, restriction of range and internal consistency

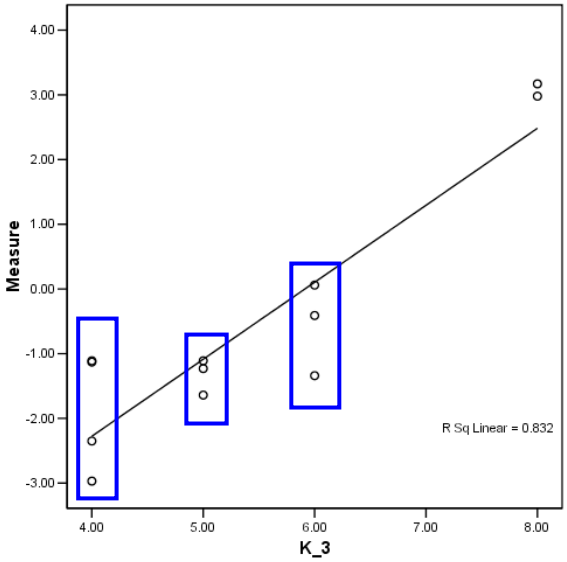
What are fit statistics and what do they mean? FACETS generates two important fit statistics: the Infit Mean-square Statistics (Infit MnSq) and the Outfit Mean-square Statistics (Outfit MnSq). Broadly, these fit statistics provide information on the consistency of ratings given by raters (and ratings received by students): whether the ratings are consistent, inconsistent, or overly consistent. In terms of rater judging behaviour, fit statistics of between 0.6 and 1.4 indicate reasonable and consistent judging behaviour (Linacre, 2003). Fit statistics that are very low (below 0.6) suggests restriction of range. This means that raters with very low mean-square fit statistics have the tendency to restrict their ratings to certain parts of the rating scale. High mean-square fit statistics, on the other hand, suggests problems of internal consistency; that is, the tendency to award the same ratings to performances of different ability level or different ratings to similar performances. This is problematic as no proper discrimination of student ability is being made.

Figure 4 shows that three raters (Raters 3, 4, and 13) display Infit and Outfit MnSq statistics of below 0.6. These low mean-square statistics suggest that these raters are over-fitting (too predictable). In other words, these raters are highly likely to display restriction of range. They tend to give similar ratings to performances of different ability level, thus not discriminating. Raters 18, 21, 11, 23, and 7, on the other hand (Figure 5), show high mean-square fit statistics. This indicates that they are not consistent (too unpredictable) in their judgment of similar performances. For performances of the same

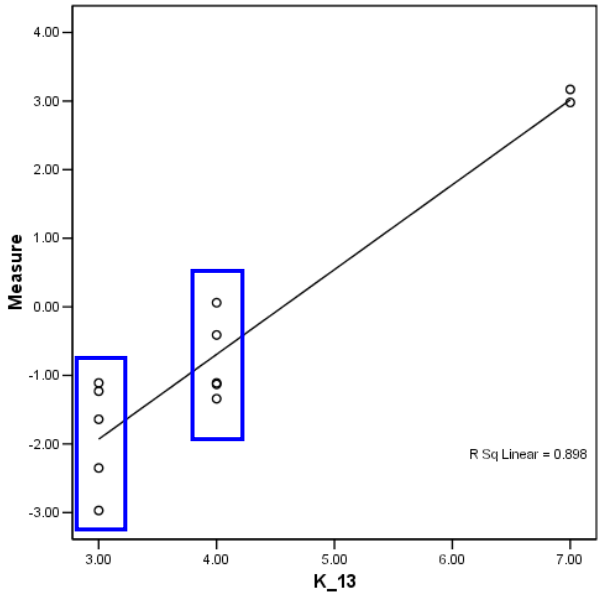
ability level, different ratings are awarded. The cross-plots in the following figures (Figures 4, 5, and 6) show what this means in terms of actual observed ratings.



Rater 4
Infit MnSq: 0.49; Outfit Mnsq: 0.47

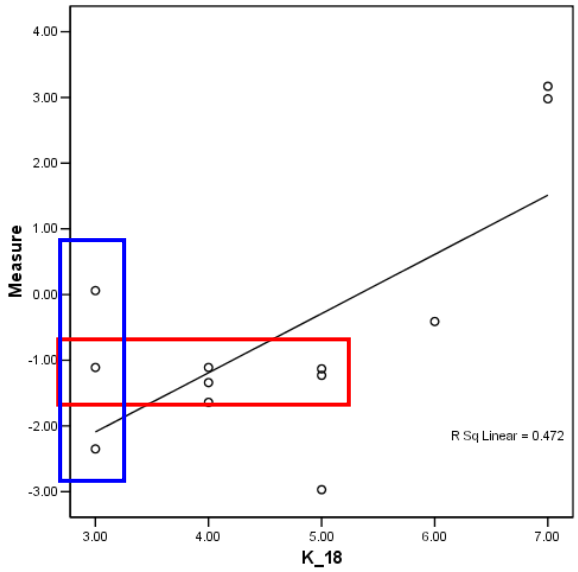


Rater 3
Infit MnSq: 0.55; Outfit Mnsq: 0.55

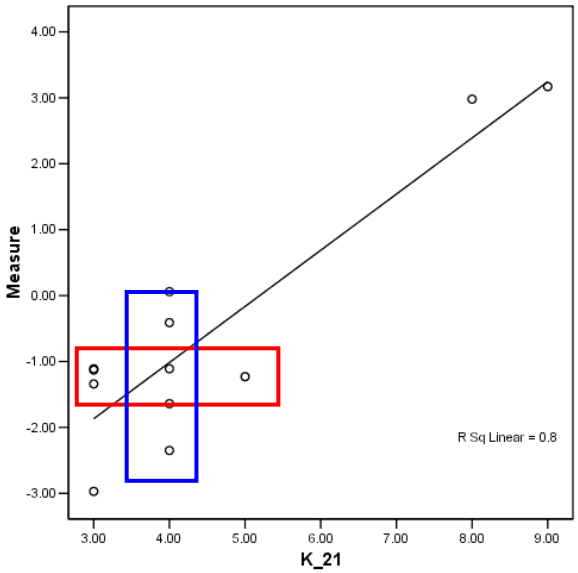


Rater 13
Infit MnSq: 0.56; Outfit Mnsq: 0.50

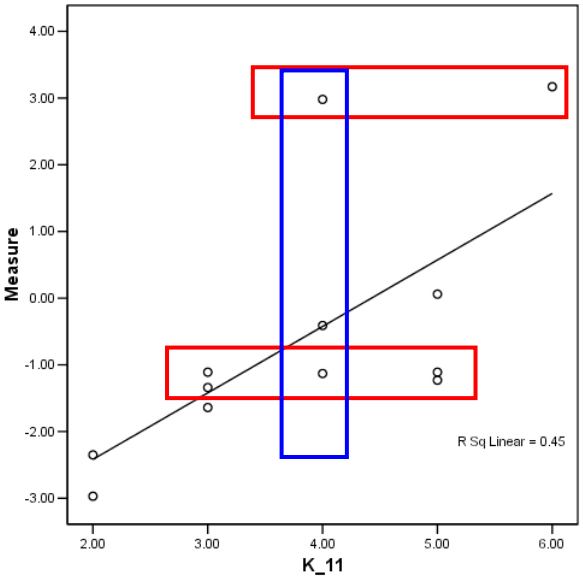
Figure 4: Cross-plots of raw ratings by raters with low mean-square fit statistics and logit measures of students. These ratings demonstrate restriction of range by raters.



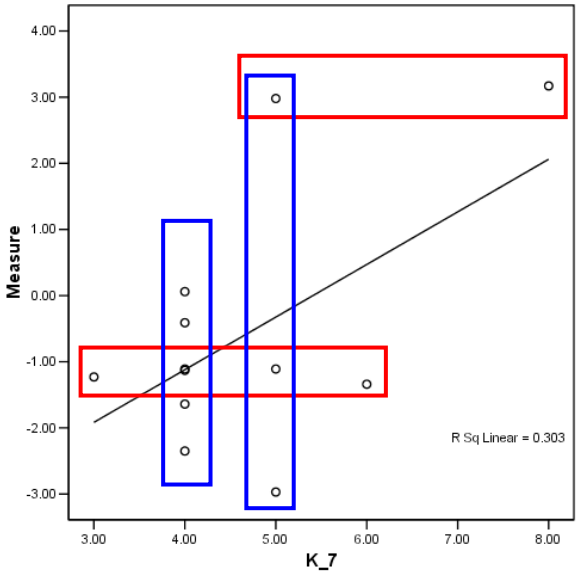
Infit MnSq: 1.76; Outfit Mnsq: 1.93
Inconsistent in judging
Rater 18 overestimates poor performance



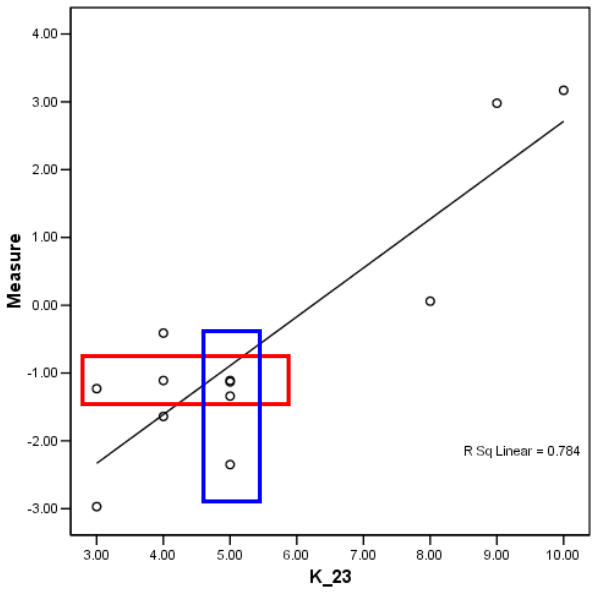
Infit MnSq: 1.80; Outfit Mnsq:1.55
Rater 21 does not discriminate performances of different ability level



Infit MnSq: 1.99; Outfit Mnsq: 2.03
Rater 11 underestimates good performance, and overestimates poor performance



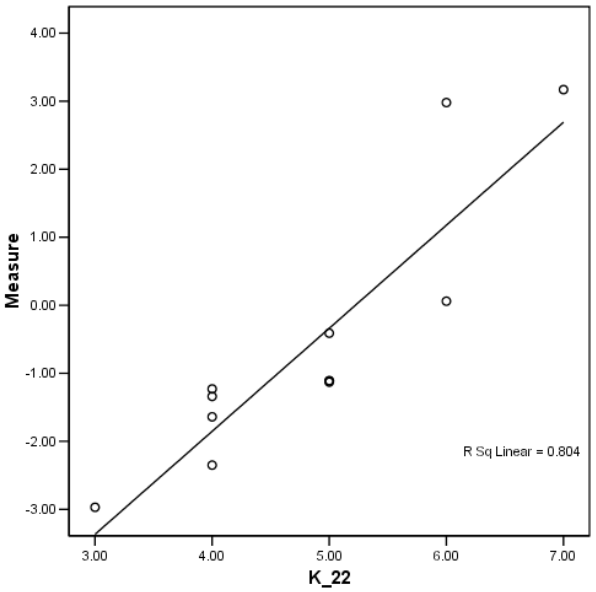
Infit MnSq: 2.19; Outfit Mnsq: 2.16
Haphazard rating. Rater 7 is unable to discriminate performances of different ability levels.



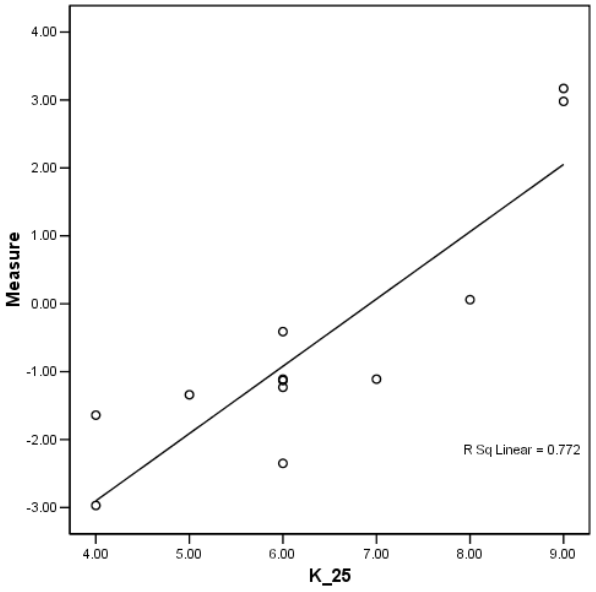
Infit MnSq: 2.09; Outfit Mnsq: 2.05
Rater 23 underestimates good performance and overestimates poor performance

Figure 5: Cross-plots of raw ratings and logit measures of raters with high fit statistics and displaying intrarater or internal inconsistency

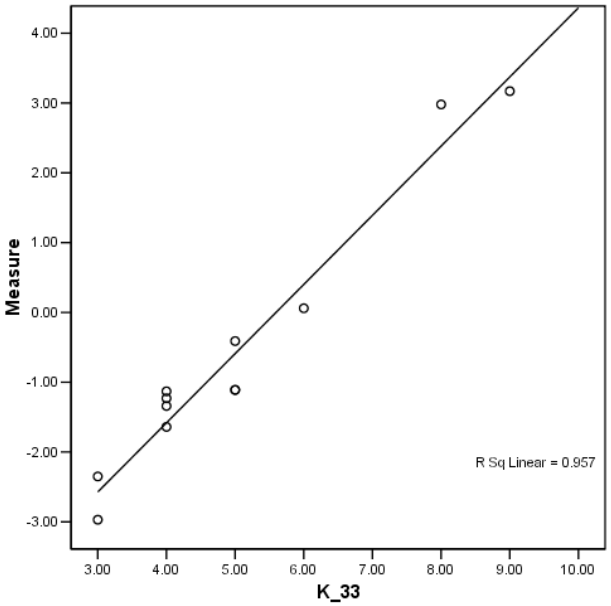
Figure 6 shows cross-plots between raw ratings and logit measures of raters with acceptable fit statistics. Notice that although these raters display some unpredictability in their judgment of similar performance, the unpredictabilities are not too severe as to degrade useful measurement, but rather confirm that the rater is rating independently, without external constraints. Also notice that Rater 33 is extremely consistent in judging the performances of students with different ability levels.



Infit MnSq: 0.76; Outfit Mnsq:0.71
Rater 22 is rather consistent in judging. Rater severity close to average measure



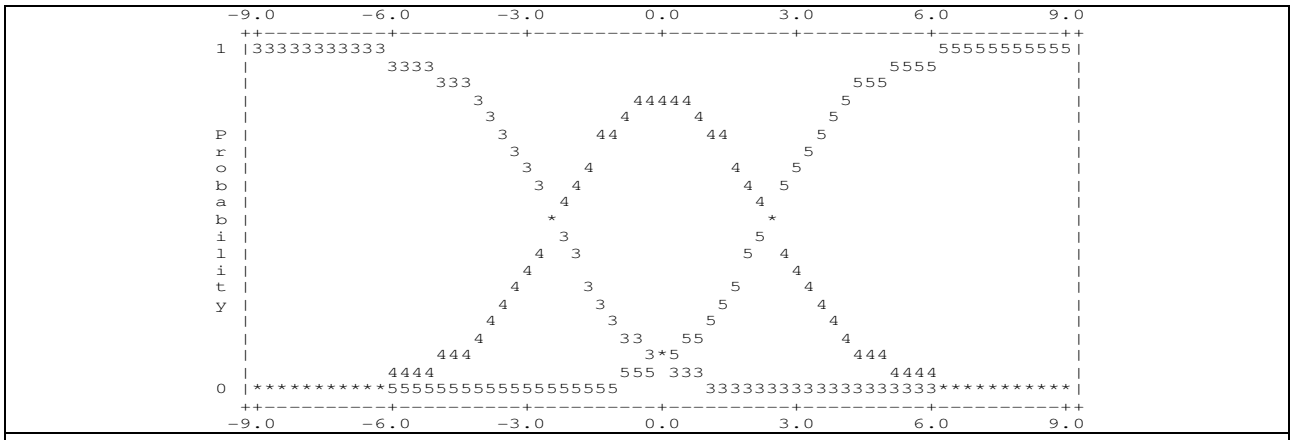
Infit MnSq: 0.89; Outfit Mnsq: 0.84
Rater 25 is the most lenient rater but rather consistent in judging performances



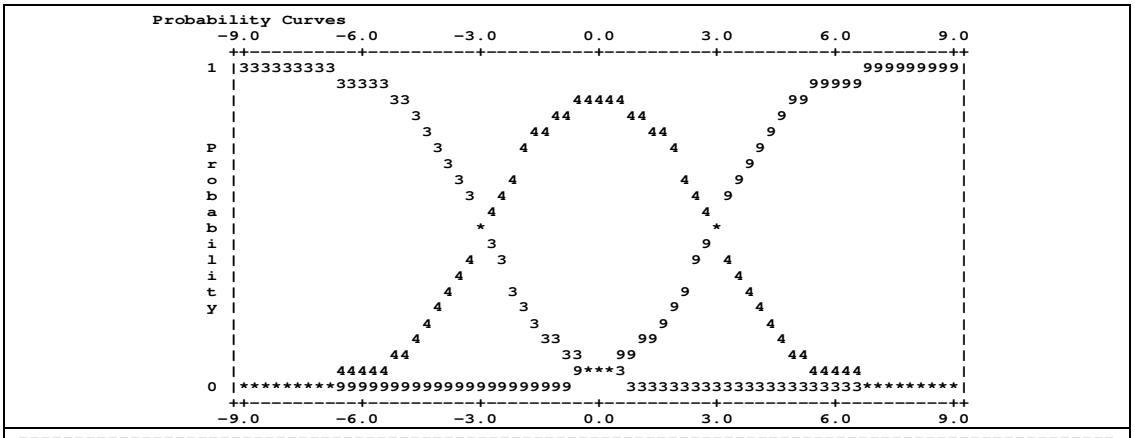
Infit MnSq: 0.74; Outfit Mnsq: 0.77
Rater 33 is very consistent in judging
Rater severity close to average measure

Figure 6: Cross-plots of raw ratings and logit measures of raters with acceptable fit statistics.

It is also possible for FACETS to model each rater to have a unique rating scale to evaluate how different raters have used the different subsets of ratings. Figures 7a and 7b are examples of raters who can only discriminate 3 performance levels (which is equivalent to three categories) and have concentrated their ratings on certain part of the rating scale (i.e., restriction of range). However, there are several raters who are able to make finer distinctions between the different levels of performances and thus provide a more accurate estimation of student ability (see Figure 8).

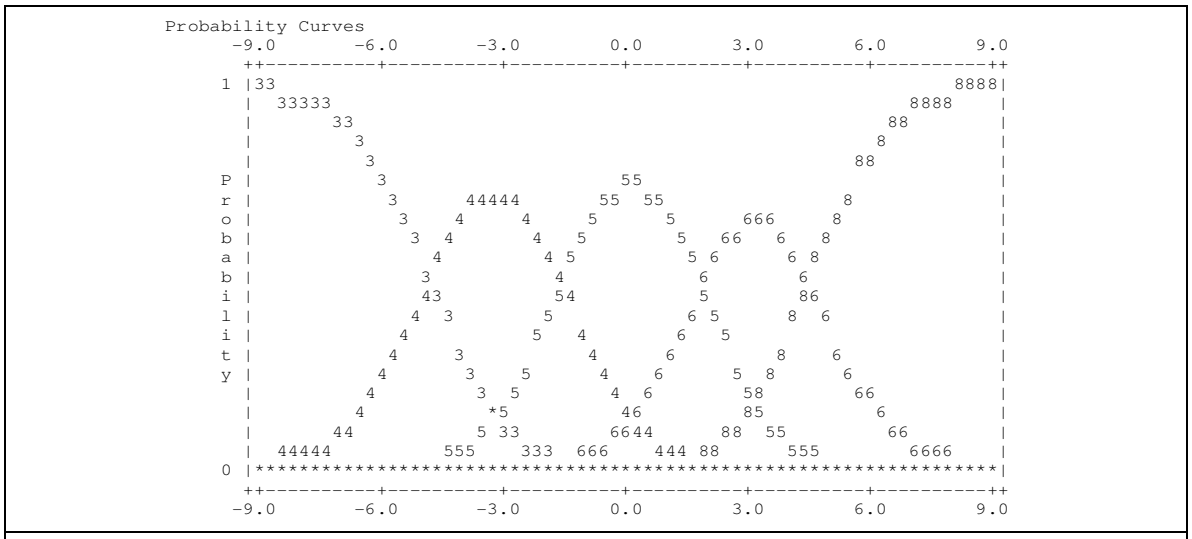


DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts	Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	CALIBRATIONS Measure	S.E.	Measure at Category	-0.5	PROBABLE from	Probabil. at	PEAK Prob
3		4	33%	-2.22	-2.42	1.0			(-3.47)		low	low	100%
4		6	50% 83%	-1.78	-1.51	1.1	-2.40	.71	.00	-2.40	-2.40	-2.41	85%
5		2	17% 100%	5.09	4.68	.1	2.40	1.98	(3.49)	2.42	2.40	2.40	100%
									(Mean)		(Modal)	(Median)	



DATA				QUALITY CONTROL			STEP		EXPECTATION		MOST	.5 Cumul.	Cat
Category	Counts	Used	Cum. %	Avge Meas	Exp. Meas	OUTFIT MnSq	CALIBRATIONS Measure	S.E.	Measure at Category	-0.5	PROBABLE from	Probabil. at	PEAK Prob
3		2	18%	-2.86	-2.05	.4			(-3.98)		low	low	100%
4		7	64% 82%	-1.11	-1.23	.3	-2.91	.86	.00	-2.90	-2.91	-2.91	90%
5													
6													
7													
8													
9		2	18% 100%	5.71	5.32	.1	2.91	2.08	(3.99)	2.90	2.91	2.90	100%
									(Mean)		(Modal)	(Median)	

Figure 7a and 7b: Example of ratings displaying restriction of range



DATA			QUALITY CONTROL			STEP		EXPECTATION	MOST	.5 Cumul.	Cat
Category	Counts	Cum.	Avge	Exp.	OUTFIT	CALIBRATIONS		Measure at	PROBABLE	Probabil.	PEAK
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category -0.5	from	at	Prob
3	1	8%	-2.46	-3.42	1.6			(-5.90)	low	low	100%
4	6	50%	-2.22	-2.60	1.6	-4.80	1.13	-3.17 -4.88	-4.80	-4.84	72%
5	3	25%	-.76	-1.81	2.8	-1.55	.75	.15 -1.52	-1.55	-1.54	73%
6	1	8%	-2.53*	4.05	8.8	1.89	2.63	3.18 1.79	1.89	1.84	64%
7											
8	1	8%	100% 4.79	4.62	.8	4.46	1.43	(5.59) 4.63	4.46	4.52	100%
						(Mean)			(Modal)	(Median)	

Figure 8: Example of rating that discriminates different levels of performances

Conclusion

As teachers, we are bound to evaluate our students' performances at one point or another. It could be an oral test, a portfolio, or a piece of writing. Hence, it is important for us to know how our judging behaviour can bring about unwanted variability or error in the measurement process and how these errors can affect the quality of ratings our students receive. We need to examine our judging behaviour and be conscious of how we rate our students' performances.

We may not be able to eliminate rater errors but we can minimize it in some ways. The use of a robust measurement model is one. Rater training is another. The most important, however, is the human factor itself. No amount of rater training can change poor attitude and no measurement model can correct for inconsistent and poor rating. If we are not willing to make the effort to ensure that our judgment of our students' performances is reliable and valid, nothing else can be done. As assessment and its procedures are at the heart of student learning (Lee King Siong, Hazita Azman, & Koo Yew Lie, 2010), matters related to valid and fair testing need to be taken seriously. It is hoped that with greater awareness of how we judge, we can be better raters and better teachers.

References

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Henning, G. (1997). Accounting for nonsystematic error in performance ratings. *Language Testing, 13*(1), 53-63.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. *Journal of Applied Psychology 63*(5)*,* 579-588.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Lee King Siong, Hazita Azman & Koo Yew Lie. (2010). Investigating the undergraduate experience of assessment in higher education. *GEMA Online™ Journal of Language Studies, 10*(1)*,* 17-33.

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

Linacre, J. M. (1998). Rating, judges and fairness. *Rasch Measurement Transactions 12*(2)*,* 630-631

Linacre, J. M. (2003). Facets (Version 3.48.0) [Computer Software and manual]. Chicago: Winsteps.com

Linacre, J. M., Engelhard, G. Jr., Tatum, D.S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research, 21*, 569-577.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing 12*(1)*,* 55-71.

Lunz. M. E. (1997). Performance examinations: *Technology for analysis and standard setting*. Paper presented at the Annual Meeting of the National Council of Measurement in Education. Chicago, IL: (ERIC Document Reproduction Service No. ED409377).

McNamara, T. F. (1996). *Measuring second language performance*. New York: Addison Wesley Longman.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing, 16*(1), 82-111.

Wigglesworth, G. (1993) Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305-336.

Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study of rater drift. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice (Volume V)* (pp. 113-134). Stamford, CT: Ablex.

About the author

Noor Lide Abu Kassim is Associate Professor at the Institute of Education, International Islamic University Malaysia with a doctorate in Psychometrics and Education Evaluation. Her doctoral research is on the setting of standards and cutoff scores using the Rasch Measurement Model. Her research interest includes testing, ESL writing, and dental education.

ISSN: 1675-8021