Exploring Rater Judging Behaviour Using the Many-Facet Rasch Model

Noor Lide Abu Kassim

Centre for Languages and Pre-University Academic Development
International Islamic University Malaysia
P.O Box, 50728
Kuala Lumpur, Malaysia

noorlide@myjaring.net/noorlide@iiu.edu.my

EXPLORING RATER JUDGING BEHAVIOUR USING THE MANY-FACET RASCH MODEL

ABSTRACT

Performance assessment, unlike the traditional fixed-response assessment, has features peculiar to its assessment setting (the task choice, the task processing conditions, the raters, the rating scale, and the rating procedures) that make it much more vulnerable to construct irrelevant variance (McNamara, 1997; Upshur & Turner, 1999). Of these potential sources of variability, those associated with raters are considered to be extensive and pose serious threats to the validity of ratings (Linacre, 1989; McNamara, 1996). For performance assessment to yield valid and reliable results, it is essential that these sources of variability are eliminated or minimized. This paper illustrates how sources of rater-related variability or rater effects can be identified and controlled for using the Many-facet Rasch Model. It also illustrates how the idiosyncratic rating behaviour of individual raters can be explicated through the use of this measurement model. In this illustration, the ratings of thirty-five English language instructors on 12 paragraphs written by new intake students at the Centre for Foundation Studies, IIUM were analysed using FACETS (Linacre, 2003), a computer application which implements the Many-facet Rasch Model. The results of the analysis indicate substantial differences in rater severity, and the presence of other rater effects.

INTRODUCTION

The advent of performance assessment has not only brought with it promises of greater validity but also risks of greater variability (McNamara, 1997). Performance assessment, unlike the traditional fixed-response assessment, has features 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, 1997; Upshur & Turner, 1999). And of these potential sources of irrelevant variance, those associated with raters have been found to be extensive, difficult to control and impossible to eliminate (Linacre, 1989; McNamara, 1996). As differences between judges' ratings and other rater effects are non-trivial and threatens the validity of test results, it is necessary that these differences are modeled, accounted for, and controlled (Linacre, 1989).

Within the Classical Test Theory (CTT), variability as a result of rater differences or effects have largely been controlled through the use of multiple raters. The reliability of ratings increases when two or more raters are utilized in the scoring procedure. Therefore, 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 a real and sufficient 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) 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 (pp. 53-54).

Secondly, the expectation that raters should be equally severe (or lenient) in their judgment cannot be supported. No two rater can be perfectly unanimous in their judgment of every performance that they encounter. The requirement within CTT that raters must agree with one another also produces counterproductive results. This is aptly argued by Linacre (1998),

> ...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 – as well as other measurement issues which are beyond the scope of this paper – there has been a shift towards the use of a more robust measurement model, as evidenced in the language testing literature. The measurement model that is now gaining popularity among language testers is the Many-facet Rasch Model, developed by Linacre (1989) to address variability that is introduced in ratings through the use of multiple raters, tasks and any other facet that constitute the testing procedure.

This paper illustrates the utility of this measurement model in adjusting for differences in rater severity in such a way that raters can be independent in their judgment of examinee performance, and at the same time, produce examinee ability estimates that are valid and reliable. It also illustrates how other sources of rater-related variability or rater effects can be identified and controlled for through the use of the model.

It is important to note that the utility of the Many-facet Rasch Model in handling rater-related variability has been discussed and explicated by other authors (e.g., Kondo-Brown, 2002; McNamara, 1996; Wigglesworth, 1993). However, these earlier papers have been rather technical and therefore, not easily accessible to the lay person. In these papers, the focus is largely on the fit statistics generated in the Many-facet Rasch analysis. How these fit statistics reflect the idiosyncratic behaviour of individual judges was not explicated for the lay person to grasp. This paper, therefore, aims to bridge that gap by elucidating the relationship between the fit statistics generated in the Many-facet Rasch analysis and what it means in terms of raters' actual judging behaviour.

Rater Effects

Research in rater judging behaviour has identified a number of rater effects that influence the validity and reliability of ratings. Chief among these is rater severity. Rater severity refers to the tendency for raters "to consistently provide ratings that are lower or higher than is warranted by examinee performances" (Engelhard, 1994). Rater severity, though considered a serious threat within CTT, is a non-issue when the Many-facet Rasch model is used as it is adjusted for in the estimation of examinee ability.

Another type of rater effect relates to the internal consistency of ratings given by individual raters (i.e., intrarater consistency). Problems of internal consistency can be seen when raters are not consistent or constant in their judgment of similar performances. The halo effect is yet another type of undesirable rater effect. A halo effect is said to be present when "a rater fails to distinguish between conceptually distinct and independent aspects of an examinee's composition" (Engelhard, 1994, p. 98). This type of rater effect can be seen when analytic-type rating scales are used.

Central tendency and restriction of range are two other types of rater effects. Central tendency occurs when middle categories are predominantly used by raters. The frequent use of middle categories reflects raters' reluctance to use extreme categories, and as these ratings lack heterogeneity, this inevitably results in overly consistent fit statistics (Engelhard, 1994). Central tendency can also be detected by examining the pattern of category usage.

Restriction of range, on the other hand, happens 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 effects 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).

The Many-Facet Rasch Model

The aim of the testing process is to provide fair and accurate estimation of examinee performance. Therefore, the measure that is given to an examinee derived from a particular rater or raters who rated that examinee must be independent of the particular rater or raters (Linacre, 1989) that were used in the judging process. This is to insure that there is consistent measurement of examinees and a valid inference of examinee ability (Lunz, 1997).

The Many-facet Rasch model (MFRM) developed by Linacre (1989) is particularly significant in this respect. MFRM 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 judge 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 s/he uses the rating scale (Linacre, 1989; Linacre et al., 1994). Interactions between facets in the testing process can also be modelled and statistically tested. In addition, MFRM is also able to detect other rater effects such as restriction of range, the halo effect and internal inconsistency through the use of 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*.

METHODOLOGY

Raters:
The raters used in this study were 34 instructors of English language at the Centre for Foundation Studies of the International Islamic University Malaysia. They were asked to participate in this study as part of a standardization exercise organized by the Testing and Measurement Unit of the English Language Department. These raters were requested to rate 12 paragraphs written by new-intake students as part of the placement test conducted by the IIUM. The academic qualifications of these instructors ranged from certificate to master's degree. These instructors had also been with the Centre for at least a year.

Materials and Method:
The 12 paragraphs were placed in random order and scored using a holistic scoring procedure. The scoring scale used in the judging of the paragraphs was a 10-point holistic rating scale, which was developed by the Testing and Measurement Unit. These paragraphs represent exemplars of writing at each band of the rating scale and they were selected by the Testing and Measurement Unit at the Centre. A complete judging plan was used in this study, where all the raters were required to rate all the writing samples. Therefore, a total of 408 ratings (34 x 12) were subjected to analysis.

Data Analysis:

The raw ratings given by each rater were analyzed using FACETS (Linacre, 2003), a computer application which implements the Many-facet Rasch Model, and SPSS version 12.0. FACETS was used to estimate examinee ability, rater severity and identification of other rater effects. SPSS, on the other hand, was used to generate descriptive statistics of the distribution of raw ratings and for plotting raters' raw ratings and examinee ability estimates derived from the FACETS analysis.

RESULTS

I. Distribution of Raw Ratings

Figure 1 shows the distribution of raw ratings for each paragraph. From the boxplots, it is evident that raters differ in the severity of their judgment of the individual paragraphs. The difference in raw ratings for the paragraphs 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 also interesting to note that although Paragraph 10 is generally seen as a good paragraph by most raters, there are several raters who had given this paragraph very low ratings. Another important observation has to do with the placement of the ratings given by raters in relation to the passing score. As the passing score is a rating of 5, it is clear that only 2 paragraphs (Paragraph 10, and 12) have been clearly judged; as clear passes, in this case. The other paragraphs have been passed by some raters but failed by others. This suggests that the use of raw ratings has serious implications on examinee classification.

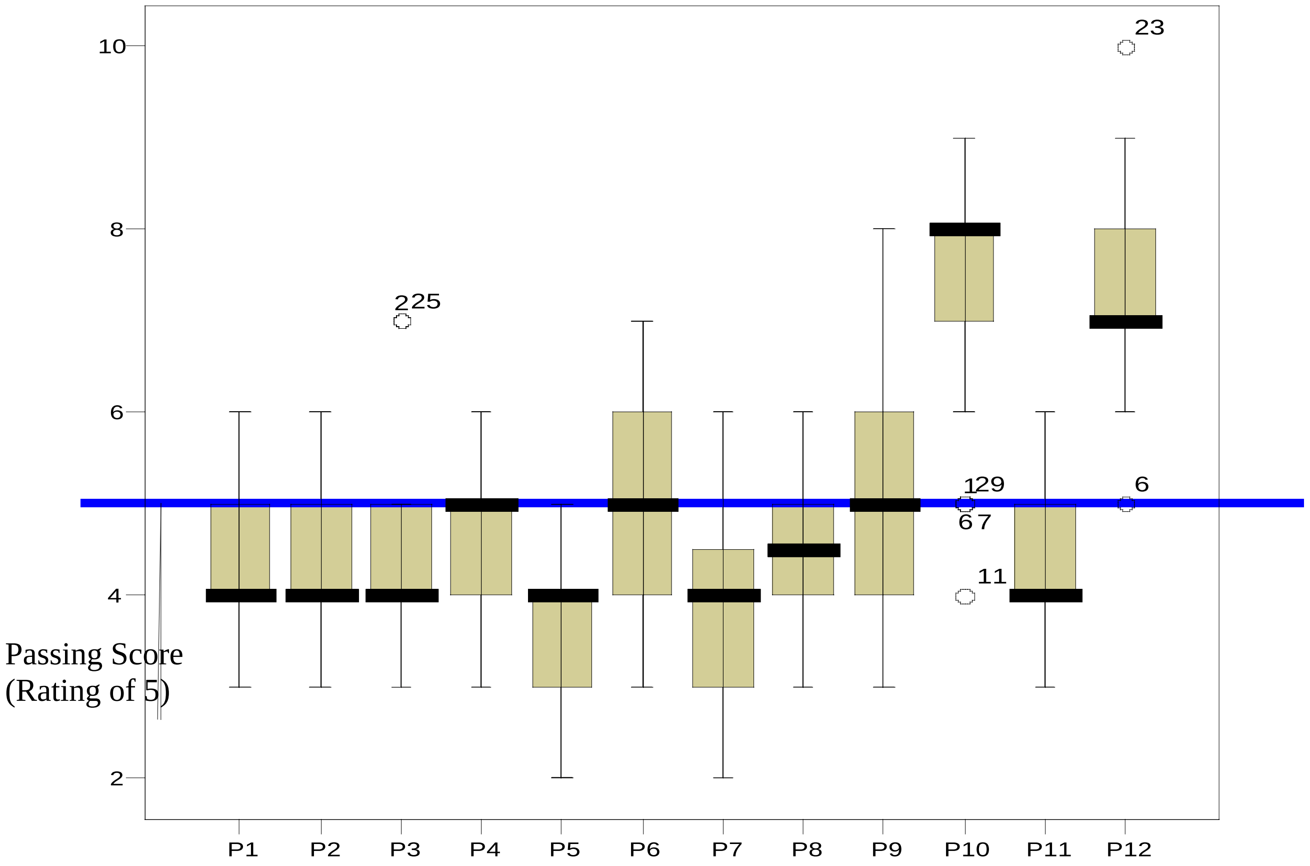


Figure 1: Distribution of Ratings for Individual Paragraphs

II. FACETS Analysis

FACETS Summary

Figure 2 gives a graphic presentation of examinee ability and rater severity generated by FACETS. The first column on the right is the logit scale, the measurement unit in which examinees and rater severity are measured. The second column gives the examinee distribution whereas the third column presents the rater distribution.

Examinees are ordered along the logit scale with the most able at the top and the least able at the bottom of the scale. In this figure, the most able examinee (Examinee/Paragraph 12) has a measure of approximately +3.2 logits, and the least able (Examinee/Paragraph 5) has a measure of approximately -3.0 logits. From the examinee distribution, it is evident that there is a considerable amount of variation of ability among examinees (a range of about 6 logits).

On the other hand, 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 examinee distribution. This indicates that these raters differ considerably in their severity level. This also suggests that examinees' performances would be either grossly underestimated or overestimated if raw ratings are used in the reporting of the results.

```
----------------------------------
|Measr|+examinee   |-reader|S.1  |
----------------------------------
+   4 +            +       +(9)  +
|     |            |       |     |
|     |            |       | 8   |
|     |            |       |     |
|     |            |       |     |
|     | 12         |       | --- |
+   3 + 10         +       +     +
|     |            |       |     |
|     |            |       |     |
|     |            |       | 7   |
|     |            |       |     |
|     |            | ***   |     |
+   2 +            +       + --- +
|     |            | *     |     |
|     |            |       |     |
|     |            | *     |     |
|     |            |       |     |
|     |            | *     | 6   |
|     |            |       |     |
+   1 +            + *     +     +
|     |            | *     |     |
|     |            | *     | --- |
|     |            | *     |     |
|     |            | **    |     |
|     |            | ***   |     |
|     |            | *     |     |
*   0 * 9          *       *     *
|     |            | **    | 5   |
|     |            |       |     |
|     | 6          | ****  |     |
|     |            | **    |     |
|     |            |       |     |
|     |            | ***   |     |
+  -1 +            + **    + --- +
|     | 1   3   4  | *     |     |
|     | 11  8      |       |     |
|     |            |       |     |
|     | 2          | *     |     |
|     |            | **    |     |
|     |            |       |     |
+  -2 +            +       +     +
|     |            |       | 4   |
|     | 7          |       |     |
|     |            | *     |     |
|     |            |       |     |
|     |            |       |     |
|     |            |       |     |
+  -3 + 5          +       +(2)  +
----------------------------------
|Measr|+examinee   | * = 1 |S.1  |
----------------------------------
```

Figure 2: Examinee Ability and Rater Severity Distributions

Fit Statistics

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 gives information on the consistency of ratings given by raters (and ratings received by examinees): whether they are consistent, inconsistent or overly consistent. In terms of rater judging behaviour, fit statistics of between 0.6 to 1.3 indicate reasonable and consistent judging behaviour. Fit statistics that are very low (below 0.6) suggests rater effect of restriction of range. This means that raters with very low fit statistics have the tendency to restrict their ratings to certain parts of the rating scale. High 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. This is problematic as no proper discrimination of examinee ability is being made. In relation to examinees, the fit statistics indicate whether an examinee has been consistently or inconsistently rated by raters. They also give indications of unexpectedly severe or lenient ratings received by the examinee.

Examinee Statistics

Table 1 presents the examinee statistics derived from the FACETS output. From the fit statistics it is evident that Examinee/Paragraph 2 is overly fitting (Infit MnSq: .57; Outfit MnSq: .53). This suggests that there is general agreement as to the ability level of this examinee/paragraph. Examinee/Paragraph 10, on the other hand, showed rather high Infit MnSq and Outfit MnSq statistics. What this means is that for this examinee/paragraph, there are some unexpected ratings that have been awarded. Examinee/Paragraph 5 also has a similar problem but to a lesser degree. Examinee/Paragraph 12, shows acceptable fit although it has two outlying ratings (Refer to Figure 1). It is important to note that MFRM, similar to other Rasch models expects some variation in ratings. And the variation found in ratings given to Examinee/Paragraph 12 is not more than what is expected by the model.

Table 1: Examinee Statistics

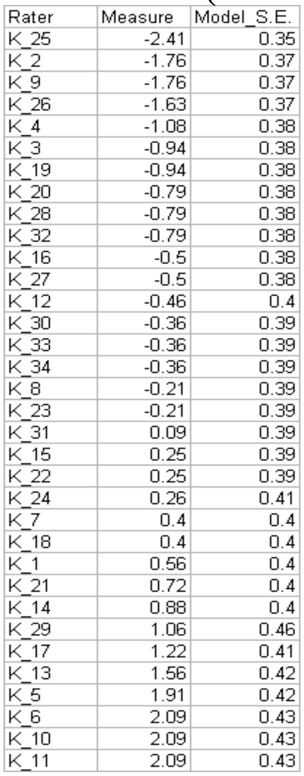
```
Table 7.1.1  examinee Measurement Report  (arranged by MN).

-------------------------------------------------------------------------------------------------
| Obsvd   Obsvd  Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.|                    |
| Score   Count Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| Nu Examinee        |
-------------------------------------------------------------------------------------------------
|    124     34     3.6   3.63|  -2.97   .27 | 1.39  1.4  1.40  1.5|  .60 |  5 5               |
|    133     34     3.9   3.89|  -2.35   .26 |  .92  -.2   .93  -.2| 1.07 |  7 7               |
|    144     34     4.2   4.20|  -1.64   .25 |  .57 -1.9   .53 -2.2| 1.45 |  2 2               |
|    149     34     4.4   4.34|  -1.34   .24 |  .74 -1.0   .74 -1.0| 1.28 | 11 11              |
|    151     34     4.4   4.39|  -1.23   .24 | 1.25   .9  1.33  1.2|  .68 |  8 8               |
|    148     33     4.5   4.44|  -1.13   .24 |  .74 -1.0   .73 -1.0| 1.23 |  1 1               |
|    153     34     4.5   4.45|  -1.11   .24 |  .82  -.6   .83  -.6| 1.18 |  3 3               |
|    153     34     4.5   4.45|  -1.11   .24 |  .70 -1.2   .72 -1.1| 1.24 |  4 4               |
|    162     33     4.9   4.82|   -.41   .23 |  .69 -1.2   .72 -1.1| 1.28 |  6 6               |
|    171     33     5.2   5.10|    .06   .22 | 1.31  1.2  1.30  1.1|  .67 |  9 9               |
|    249     34     7.3   7.39|   2.98   .20 | 1.55  2.1  1.49  1.8|  .38 | 10 10              |
|    247     33     7.5   7.54|   3.17   .21 | 1.00   .0   .94  -.1|  .95 | 12 12              |
-------------------------------------------------------------------------------------------------
```

Rater Statistics and Rater Severity

Congruent with the spread of rater severity in Figure 1, the separation index of 2.72 and chi-square value of 284.7 significant at p<0.01 indicate that raters consistently differ from one another in overall severity. The number of exact agreements is 1913 (29.0%) out of a total of 6600 rater agreement opportunities. The most severe raters are Rater 6, 10 and 11 (+2.09 logits) while the most lenient is Rater 25 (-2.41 logits) (Table 2a). Table 2b shows the ordering of raters according to the Infit and Outfit MnSq statistics. The following tables, on the other hand, demonstrate how differences in rater severity has affected examinee performance. Table 3 gives the median ratings awarded by raters and the logit measures derived from the FACETS analysis. Table 4 gives the ranking of the examinees/ paragraphs based on the median ratings and the logit measures. Before adjustments were made to differences in rater severity (i.e. based on median rating) Examinee/Paragraph 10 was ranked first; but after adjusting for rater severity, Examinee/Paragraph 12 was ranked first (Refer to

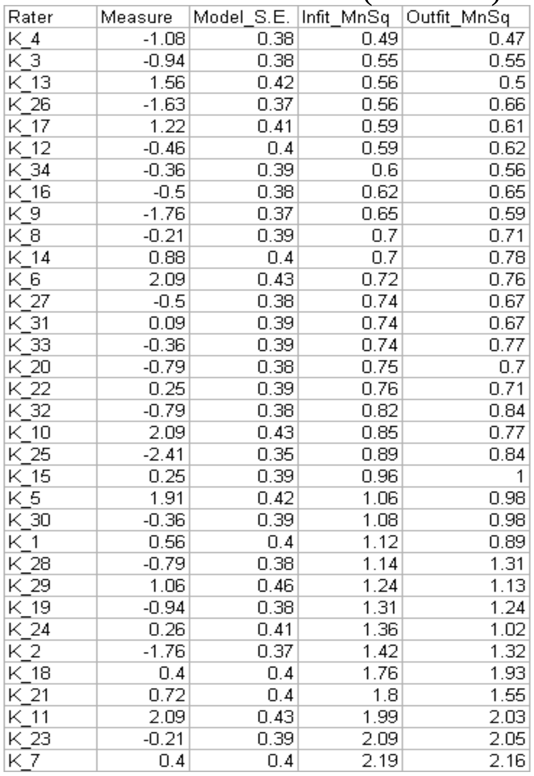
Table 4). Notice also that median ratings are unable to discriminate between performances of different ability unlike the logit measures which have been adjusted for differences in rater severity.

Table 2a: Rater Statistics(Measure Order)



Rater	Measure	Model_S.E.
K_25	-2.41	0.35
K_2	-1.76	0.37
K_9	-1.76	0.37
K_26	-1.63	0.37
K_4	-1.08	0.38
K_3	-0.94	0.38
K_19	-0.94	0.38
K_20	-0.79	0.38
K_28	-0.79	0.38
K_32	-0.79	0.38
K_16	-0.5	0.38
K_27	-0.5	0.38
K_12	-0.46	0.4
K_30	-0.36	0.39
K_33	-0.36	0.39
K_34	-0.36	0.39
K_8	-0.21	0.39
K_23	-0.21	0.39
K_31	0.09	0.39
K_15	0.25	0.39
K_22	0.25	0.39
K_24	0.26	0.41
K_7	0.4	0.4
K_18	0.4	0.4
K_1	0.56	0.4
K_21	0.72	0.4
K_14	0.88	0.4
K_29	1.06	0.46
K_17	1.22	0.41
K_13	1.56	0.42
K_5	1.91	0.42
K_6	2.09	0.43
K_10	2.09	0.43
K_11	2.09	0.43

Table 2b: Rater Statistics (Fit Order)



Rater	Measure	Model_S.E.	Infit_MnSq	Outfit_MnSq
K_4	-1.08	0.38	0.49	0.47
K_3	-0.94	0.38	0.55	0.55
K_13	1.56	0.42	0.56	0.5
K_26	-1.63	0.37	0.56	0.66
K_17	1.22	0.41	0.59	0.61
K_12	-0.46	0.4	0.59	0.62
K_34	-0.36	0.39	0.6	0.56
K_16	-0.5	0.38	0.62	0.65
K_9	-1.76	0.37	0.65	0.59
K_8	-0.21	0.39	0.7	0.71
K_14	0.88	0.4	0.7	0.78
K_6	2.09	0.43	0.72	0.76
K_27	-0.5	0.38	0.74	0.67
K_31	0.09	0.39	0.74	0.67
K_33	-0.36	0.39	0.74	0.77
K_20	-0.79	0.38	0.75	0.7
K_22	0.25	0.39	0.76	0.71
K_32	-0.79	0.38	0.82	0.84
K_10	2.09	0.43	0.85	0.77
K_25	-2.41	0.35	0.89	0.84
K_15	0.25	0.39	0.96	1
K_5	1.91	0.42	1.06	0.98
K_30	-0.36	0.39	1.08	0.98
K_1	0.56	0.4	1.12	0.89
K_28	-0.79	0.38	1.14	1.31
K_29	1.06	0.46	1.24	1.13
K_19	-0.94	0.38	1.31	1.24
K_24	0.26	0.41	1.36	1.02
K_2	-1.76	0.37	1.42	1.32
K_18	0.4	0.4	1.76	1.93
K_21	0.72	0.4	1.8	1.55
K_11	2.09	0.43	1.99	2.03
K_23	-0.21	0.39	2.09	2.05
K_7	0.4	0.4	2.19	2.16

Table 3: Comparisons between Examinee/Paragraph Median Rating and Logit Measures

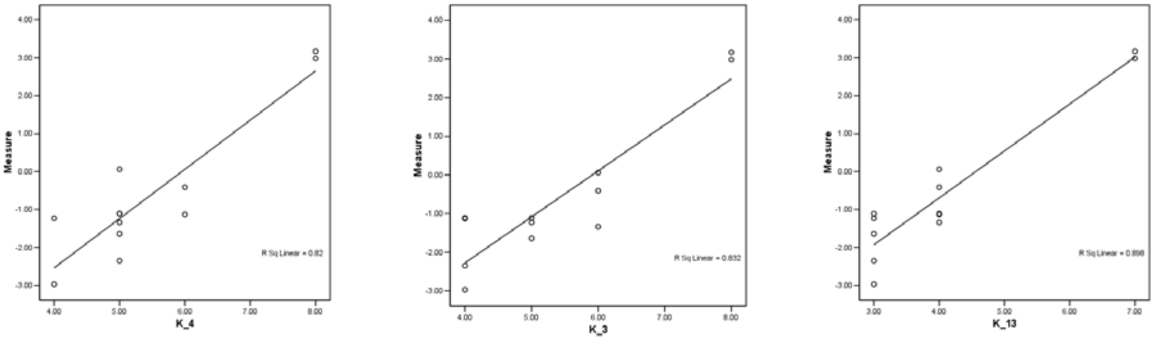
	Paragraph											
	1	2	3	4	5	6	7	8	9	10	11	12
Median Rating	4	4	4	5	4	5	4	4.5	5	8	4	7
Rasch Measure	-1.13	-1.64	-1.11	-1.11	-2.97	-.41	-2.35	-1.23	.06	2.98	-1.34	3.17

Table 4: Comparisons between Examinee/Paragraph Ranking Based on Median Rating and 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	6	9	5	5	11	4	10	7	3	2	8	1

<u>Restriction of Range and Intrarater Inconsistency:</u>

Table 2b shows that three raters (Raters 4, 3 and 13) display Infit and Outfit MnSq statistics of below 0.6. These low mean-square statistics suggest that these raters are over-fitting. In other words, these raters are highly likely to display restriction of range. Rater 18, 21, 11, 23, and 7, on the other hand, show high fit statistics. This suggests that they are not consistent in their judgment of similar performances. The cross-plots in the following figures (Figures 3, 4, and 5) show what this means in terms of actual judging behaviour.

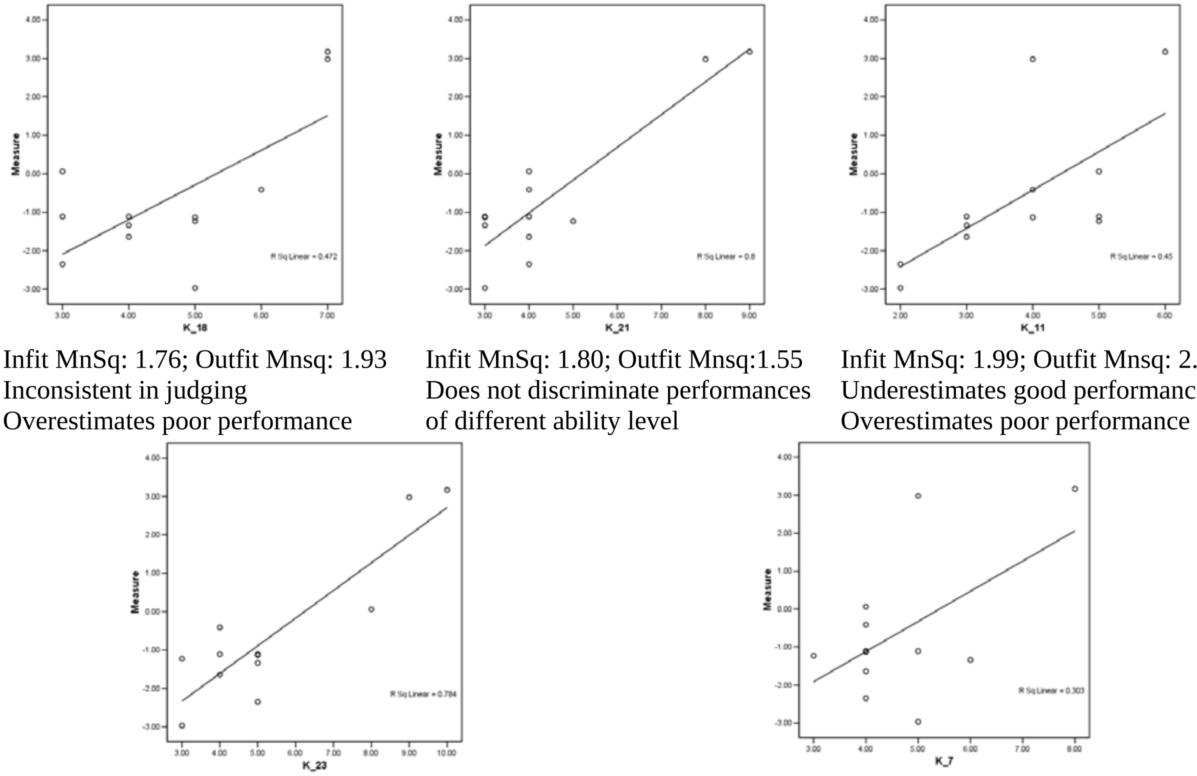


Infit MnSq: 0.49; Outfit Mnsq: 0.47

Infit MnSq: 0.55; Outfit Mnsq: 0.55

Infit MnSq: 0.56; Outfit Mnsq: 0.50

Figure 3: Cross-plots of raw ratings and logit measures of raters with low fit statistics (i.e., displaying restriction of range) .



Infit MnSq: 1.76; Outfit Mnsq: 1.93
Inconsistent in judging
Overestimates poor performance

Infit MnSq: 1.80; Outfit Mnsq:1.55
Does not discriminate performances of different ability level

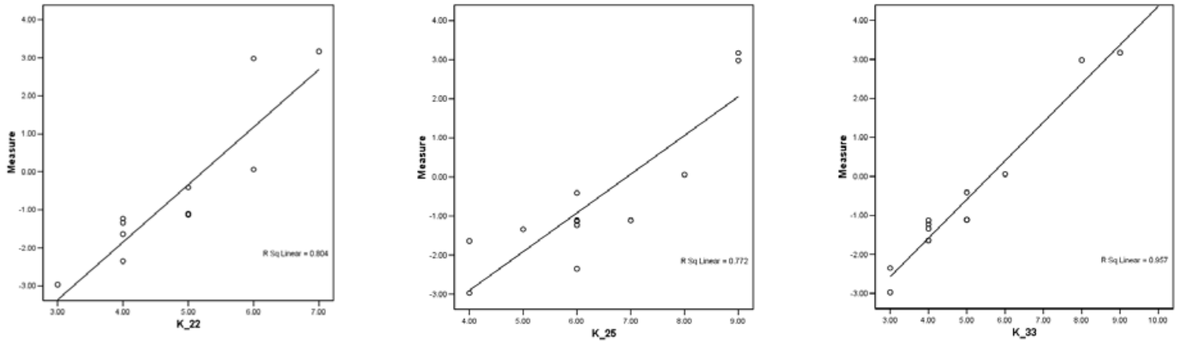
Infit MnSq: 1.99; Outfit Mnsq: 2.03
Underestimates good performance
Overestimates poor performance

Infit MnSq: 2.09; Outfit Mnsq: 2.05
Underestimates good performance and overestimates poor performance

Infit MnSq: 2.19; Outfit Mnsq: 2.16
Haphazard rating. Unable to discriminate performances of different ability levels.

Figure 4: Cross-plots of raw ratings and logit measures of raters with high fit statistics (i.e., displaying intrarater inconsistency).

The following figure shows cross-plots between raw ratings and logit measures of raters with acceptable fit statistics. Notice that although these raters display some inconsistency in their judgment of similar performance, the inconsistencies are not too severe as to degrade useful measurement. Also notice that Rater 33 is extremely consistent in judging the performances of different abilities.



Infit MnSq: 0.76; Outfit Mnsq:0.71
Rather consistent in judging
Rater severity close to average measure

Infit MnSq: 0.89; Outfit Mnsq: 0.84
Most lenient rater but rather consistent in judging performances

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

Figure 5: Cross-plots of raw ratings and logit measures of raters with acceptable fit statistics.

CONCLUSION

This paper sought to illustrate the utility of the Many-facet Rasch model in dealing with differences in rater severity as well as in identifying the presence of other rater effects in terms that can be easily understood. It is also hoped that this paper has helped readers see the usefulness of the Many-facet Rasch model in making our assessment of students' performances a fair and equitable one.

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.

Kondo-Brown, K. (2002). A FACETS Analysis of Rater Bias in Measuring Japanese Second Language Writing Performance. *Language Testing*, 19(1), 3-31.

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

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

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.