BAR99789

HECS LOTTO: DOES MARKER VARIABILITY MAKE EXAMINATIONS A LOTTERY?

Steven Barrett

Division of Business and Enterprise

University of South Australia

Abstract

Focus groups that have been conduced with undergraduate students of the Division of Business and Management at the University of South Australia revealed general concerns about marker variability and the possible impact on examination results and student performance. This study has two aims. First, to analyse the relationships between student performance on an essay style examination, the questions answered and the markers. Second, to identify and determine the nature and the extent of the marking errors on the examination.

These relationships were analysed using two commercially available software packages, RUMM and Conquest to develop Rasch Test Models. The analyses revealed minor differences in item difficulty, but considerable inter-rater variability. Furthermore, intra-rater variability was even more pronounced. Four of the five common marking errors were also identified.

Introduction

The Division of Business and Enterprise at the University of South Australia offers over 40 courses to about 6,000 undergraduate students. However, increasingly scarce teaching resources are contributing to further increases in the casualisation of teaching. One of the major responses to this situation is the introduction of the 'faculty core' of eight subjects. The introduction of these core subjects provides the Division with a vehicle through which it can attempt to realise economies of scale in teaching. These subjects have enrolments of up to 1,500 students in a semester and are commonly taught by one, sometimes two lecturers, supported by a large team of sessional tutors.

These responses may allow the Division to address some of the problems associated with its resource constraints, but they also introduce a set of other problems. Focus groups conducted with students of the Division constantly raise a number of issues that concern the students of the Division. Three of the more important issues identified at these meetings are;

- consistency between examination markers (inter-rater variability);
- consistency within examination markers (intra-rater variability); and

- differences in the difficulty of examination questions (inter-item variability).

The students argue that if there is significant inter-rater variability, intra-rater variability and inter-item variability then student examination performance becomes a function of the marker and questions, rather than the previous semester's teaching and learning experiences.

Design of the study

The aim of the project is to use Latent Trait Theory, by employing the Rasch Model, to determine whether student performance in essay examinations is a function of the person who marks the examination papers and the questions students attempt, rather than an outcome of the preceding semester's learning experiences. The study investigates the following four questions:

1. To what extent does the difficulty of items in an essay examination differ?
2. What is the extent of inter-rater variability?
3. What is the extent of intra-rater variability?
4. To what extent are the five rating errors present?

The project analyses the results of the semester 1, 1997 final examinations results in Communication and the Media. The 833 students who sat this examination were asked to answer any four questions from a choice of 12. The answers were arranged in tutor order and eight tutors marked all of the papers written by their students. The unrestricted choice on the paper and the decision to allow tutors to mark all questions answered by their students maximises the crossover between items. However, the raters did not mark answers written by students from other tutorial groups. Hence, the relationship between the rater and the students cannot be separated. It was therefore decided to have all of the tutors mark a random sample of all of the other tutorial groups in order to facilitate the separation of raters, students and items. In all 19.4 per cent of the papers were double marked. The 164 double marked papers were then analysed separately in order to provide some insights into the effects of student performance by fully separating raters, items and students.

The Rasch Model

The examination results were analysed using a Rasch Model, which is a latent trait model based on the premise that the performance of students can be determined by an underlying or latent trait that is not observable. The latent trait that enables students to correctly answer an examination question is usually referred to ability. The model is probabilistic and is concerned with defining and predicting the probability of obtaining a correct answer to question as a function of the underlying trait, in this case student ability. The model attempts to develop and specify the relationship between the observable performance of students and the unobservable latent trait that underlies their performance.

The Rasch Model predicts the odds or probability of a student obtaining a correct answer on a question in terms of two parameters, one relating to the difficulty of the items on the test and the other to the ability of the students. The basis of the model is that the relationship between item difficulty and student ability determines the performance of students on a test. That is, a student with greater ability should also have a higher chance of success on a particular question than a less able person. Conversely, a person of any level of ability would have a greater chance of success on a less difficult question than on a more difficult question. The probability of the success of a student on a question can be specified as a function of the difference between the ability of the student and the difficulty of a question, where both ability and difficulty are measured on the same linear scale using the same units

of measurement. The Rasch Model facilitates the construction of an interval scale that allows two disparate concepts, such as student ability and item difficulty to be measured and compared.

The relationship between ability and difficulty can be expressed simply in the forms of odds. The odds of getting a question correct (O) is equal to the ability of the student (A) multiplied by the easiness of the question (E), that is, O = AE. To illustrate this relationship, if ability was zero then all questions would be impossible. On the other hand, if ability were very high, then all but the most difficult questions would be easy. At the mid-point between these extremes of student ability and easiness of the question are reciprocals such that AE = 1. As this is a probabilistic model the range of both ability and easiness is from zero to one. Hence, the student and the question are perfectly matched and the odds of success would be 1:1 and the probability of success would be 0.5.

If logs of both sides of the relationship are taken, then an additive form of the equation is obtained, logObserved = logAbility + logEasiness. The difficulty of the question can now be substituted for easiness. The use of a difficulty rather than an easiness parameter means that the odds of a correct answer falls as the difficulty of the question rises. This gives, logObserved = logAbility - logDifficulty. The Rasch Test Model can now be expressed in probabilistic terms;

$$\text{Probability}_{(\text{correct response})} = \frac{1\log(A - D)}{1 + 1\log(A - D)}$$

Which in turn can be expressed as an exponential function;

$$\text{Probability}_{(\text{correct response})} = \frac{e\text{^}(\text{A} - \text{D})}{1 + e\text{^}(A - D)}$$

This is the expression for the probability of a single student getting one question correct. The expression is then summed across all questions and all students in order to develop the model. Finally the performance of students can be explained in terms of only two parameters, one for the student (ability) and one for the item (difficulty). More importantly, both parameters are expressed in the same units (logits) and measured on the same scale.

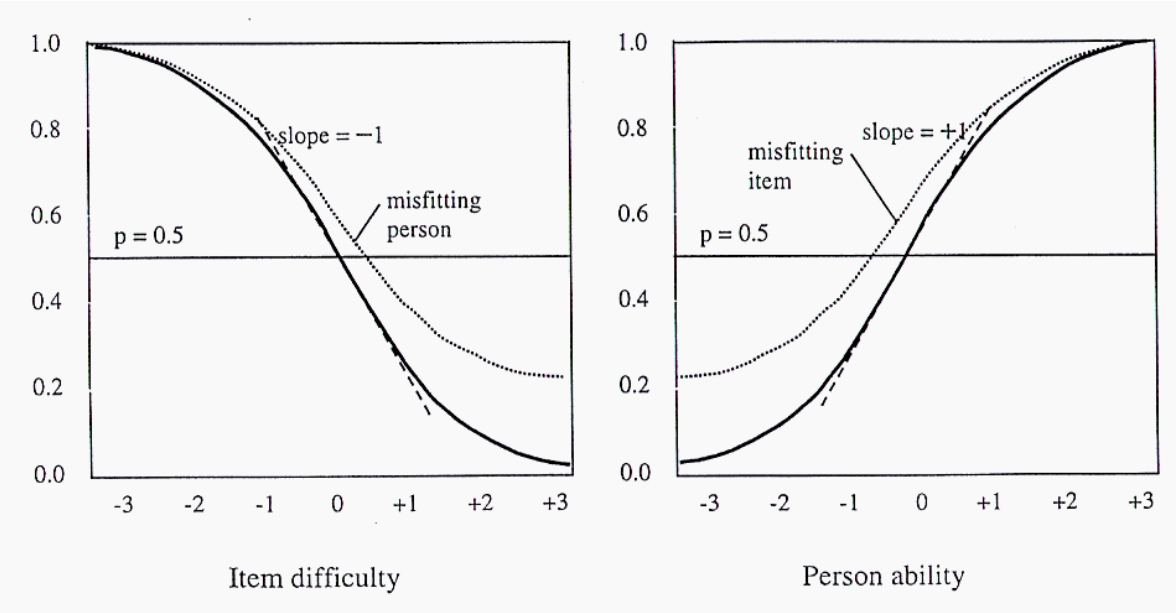
Calculation of the Rasch Model involves the collection of the responses to a set of test items, and an estimation of the values in the parameters in the model for the items and the students that best fit the data. Iterative computer procedures are used to calculate the maximum likelihood estimates of the parameters. Initial estimates are made for item difficulties based on the number of correct answers. Then initial estimates are made for student ability based on their scores. The initial estimates of ability are then used to improve the estimates of item difficulty, which in turn are used to improve the estimates of student ability. The process is iterated in order to maximise the fit of the parameter estimates to the test data.

Five ratings errors

Previous research into performance appraisal has identified five major categories of rating errors; severity or leniency, the halo effect, the central tendency effect, restriction of range and inter-rater reliability or agreement (Saal et al 1980). Engelhard and Stone (1998) have demonstrated that the statistics obtained from the Rasch Model can be used to measure

these five types of error. This section briefly outlines these errors in rating and identifies the underlying questions that motivate concern about each type of error. The discussion describes how each type of rating error can be detected by analysing the statistics obtained after developing a Rasch Model. The present study extends this procedure by demonstrating how Item and Person Characteristic Curves can also be used to identify these rating errors.

Figure 1: Item and Person Characteristic Curves



Source: Keeves and Alagumalai 1999, 30.

Rater severity or leniency

Rater severity or leniency refers to the general tendency on the part of raters to rate consistently students higher or lower than is warranted on the basis of their responses (Saal et al 1980). The underlying questions that are addressed by indices of rater severity focus on whether there are statistically significant differences in rater judgments. The statistical significance of rater variability can be analysed by examining the rater estimates (Tables 3 and 5). The estimates for each rater should be compared with the expert in the field, that is the subject coordinator in this instance. If the Person Characteristic Curve for a particular rater lies to the right of the expert's then that rater is more severe. On the other hand, a Person Characteristic Curve lying to the left implies that the rater is more lenient.

The halo effect

The halo effect appears when a rater fails to distinguish between conceptually distinct and independent aspects of student performance (Thorndike 1920). For example, if the rater does not distinguish between essential and non-essential content. Then the rater simply treats all the content as if it were either essential or non-essential. That is, the rater is taking a holistic approach to the paper and is ignoring any other domains or criteria that the item has been constructed to measure (Engelhard 1994). Evidence of a halo effect can be obtained from a Rasch Test Model by examining the mean square error statistics, or weighted fit MNSQ in Tables 3 and 5. If these statistics are very low, that is less than 0.6, then raters may not be rating items independently of each other (Engelhard and Stone 1998). The shape of the Person Characteristic Curve demonstrates the presence or absence of the halo effect. A flat curve, with a vertical intercept significantly less than 1.00 or

which is tending towards a value significantly greater than zero as item difficulty rises, is an indication of the halo effect.

Central tendency

Central tendency describes situations in which the ratings are clustered around the mid-point of the rating scale. This reflects reluctance by raters to use the extreme ends of the rating scale. This is particularly problematic when using a polycotomous rating scale, such as the one used in this study. This error can simply be detected by examining the range of marks of each rater. In addition, the mean square error statistics will be lower than expected in Tables 3 and 5 (Engelhard and Stone 1998). Central tendency can also be seen in the Item Characteristic Curves, especially if the highest ability students consistently fail to attain a score of 1.00 on the vertical axis and the vertical intercept is significantly greater than zero.

Restriction of range

Restriction of range is related to central tendency, but it is also a measure of the extent to which the obtained ratings discriminate among different students with respect to their different performance levels (Engelhard 1994). The underlying question that is addressed by restriction of range indexes focus on whether there is a statistical significance in item difficulty as shown by the rater estimates. Significant differences in these indices demonstrate that raters are discriminating between the items. The amount of spread also provides evidence relating to how the underlying trait has been defined. This is shown if the weighted fit MNSQ statistic for the item is greater than 1.30 or less than 0.77 (Engelhard and Stone 1998). These relationships are also reflected in the shape of the Item Characteristic Curve. If the weighted fit MNSQ statistic is less than 0.77, then this curve will have a very steep upward sloping section, demonstrating that the item discriminates between students in a very narrow ability range. On the other hand, if the MNSQ statistic is greater than 1.30, then the curve will be very flat with little of no steep middle section to give the characteristic "S" shape. Such an item fails to discriminate effectively between students.

Inter-rater reliability or agreement

Inter-rater reliability or agreement is based on the concept that ratings are of a higher quality if two or more independent raters arrive at the same rating. In essence, rating error reflects a concern with consensual or convergent validity. The model fit statistics obtained from the Rasch Model provides evidence of this type error (Engelhard and Stone 1998). It is unrealistic to expect perfect agreement between a group of raters. Nevertheless, it is not unrealistic to seek to obtain consistent ratings from raters. Indications of this type of error can be obtained by examining the mean square errors for both raters and items. Lower values reflect more consistency or agreement or a higher quality of ratings. Higher values reflect less consistency or agreement or a lower quality of ratings. Ideally these values should be 1.00 for the unweighted MNSQ and 0.00 for the unweighted T statistic. Mean square values greater than 1.5 suggest that raters are not rating items in the same order. The unweighted MNSQ statistic is the slope at the point of inflection of the Person Characteristic Curve. Ideally this slope should be negative 1.00. Increased deviation of the slope from this value implies less consistent and less reliable ratings.

Table 1: Summary Table

Rater error	**Features of the curves if rater error present**	**Features of the statistics if rater error present**
Leniency	Need to compare Person Characteristic Curve with the 'experts'	Rater estimates; • Check estimate of leniency; • Lower error term implies more consistency
Halo effect	Person Characteristic Curve • Vertical intercept much less than 1; • Does not tend to 0 as item difficulty rises.	Rater estimates; • Weighted fit MNSQ < 1
Central tendency	Item Characteristic Curve • Vertical intercept much greater than 0; • Does not tend to 1 as student ability rises.	Item estimates; • Unweighted fit MNSQ >> 1; • Unweighted fit T >> 0.
Restriction of range	Item Characteristic Curve • Steep section of curve occurs over a narrow range of student ability; or • Curve is very flat with no distinct "S" shape.	Item estimates; • Weighted fit 0.77 <MNSQ < 1.30
Relaibility	Person Characteristic Curve; slope at point of inflection significantly greater than or less than 1.00.	Rater estimates; • Weighted fit MNSQ >> 1; • Weighted fit T >> 0.

Phase one of the study: Initial questions

At present, the analysis of examination results and student performance at the University of South Australia is not very sophisticated. The analysis is usually confined to an examination of the range and other measures of central tendency using a spreadsheet program such as Excel. Staff with appropriate skills in educational measurement are pretty rare in most university business faculties. Hence, sophisticated analyses of student performances are only very occasionally conducted.

Table 2: Average Raw Scores for each Question for all Raters.

Rater	Question number												Mean score	n[1]
	1	2	3	4	5	6	7	8	9	10	11	12		
1	7.1	7.0	6.8	7.0	7.2	7.4	7.0	7.2	7.0	7.3	7.5	7.1	28.4	26
2	6.6	6.2	6.5	6.8	6.7	7.2	6.7	6.9	6.8	6.8	6.5	6.8	26.8	225
3	7.2	6.7	6.4	7.3	7.0	8.0	6.1	6.5	7.2	6.1	6.0	5.9	26.5	71
4	7.2	7.1	6.9	7.3	7.6	7.3	7.2	7.0	7.0	6.9	7.0	7.2	28.6	129
5	5.4	6.4	6.0	5.5	6.0	6.5	5.8	5.8	6.7	5.6	5.7	5.9	23.8	72
6	7.1	7.1	6.8	6.8	7.7	7.7	7.3	7.6	7.3	7.2	6.8	7.6	29.1	161
7	6.5	6.8	6.5	6.7	7.4	6.5	6.6	8.0	7.9	7.4	6.9	7.3	28.5	70
8	6.6	6.4	6.5	6.5	7.2	7.0	6.8	7.0	7.2	6.9	6.6	6.9	27.8	79
All	6.8	6.5	6.5	6.7	7.1	7.2	6.8	6.9	7.0	6.8	6.6	6.8	27.4	833

Note 1: n signifies the number of papers marked by each tutor, N = 833.

An Excel analysis for these data constitutes phase one of this study. Such an analysis reveals some interesting differences that should raise some interesting questions for the subject coordinator to consider as part of her curriculum development process. Table 2 shows considerable differences in question difficulty and the leniency of markers. Rater 5 is the hardest and Rater 6 the easiest, while Item 6 appears to be the easiest and Items 2 and 3 the hardest. But are these the correct conclusions to be drawn from these results?

Phase two of the study

A more sophisticated analysis of these results using Rasch Scaling, phase two of the study, tells a very different story. An analysis of the raters (Table 3) and the items (Table 4) using Conquest provides a totally different set of insights into the performance of both raters and items. Table 3 reveals that Rater 1 is the most lenient marker, not Rater 6, with the minimum

estimate value. He is also he most variable, with the maximum error value. Indeed, he is so inconsistent that he does not fit the Rasch Model, the unweighted fit MNSQ is significantly different from 1 and the T statistic is greater than 2. Nor does he discriminate well between students, as shown by the maximum value for the weighted fit MNSQ statistic, which is significantly greater than 1.30. The subject coordinator is Rater 2 and this table clearly shows that she is an expert in her field who sets the appropriate standard. Her estimate is the second highest, so she is setting a high standard. She has the lowest error statistic, which is very close to zero, so she is the most consistent. Her unweighted fit MNSQ is very close to 1.00 while her T statistic is closest to 0.00. She is also the best rater when it comes to discriminating between students of different ability as shown by her weighted fit MNSQ statistic is one of the few in the range 0.77 to 1.30.

Table 3 confirms that Rater 5 is now the toughest marker, while Rater 6 is appearing as one of the better, more consistent markers who fits the Rasch Model well, especially with respect to his ability to discriminate between students.

Table 3: Raters, Summary Statistics

VARIABLES UNWGHTED FIT WGHTED FIT

---

rater LENIENCY ERROR MNSQ T MNSQ T

---

1 -0.553 0.034 1.85 2.1 1.64 3.8

2 0.159 0.015 0.96 -0.1 0.90 -1.3

3 0.136 0.024 1.36 1.2 1.30 2.5

4 -0.220 0.028 1.21 0.9 1.37 3.2

5 0.209 0.022 1.64 2.0 1.62 4.8

6 0.113 0.016 1.29 1.3 1.23 2.3

7 0.031 0.024 1.62 1.9 1.60 4.6

8 0.124*

N= 833

Table 4 summarises the item statistics that were obtained from Conquest. The results of this table also do not correspond well to the results presented in Table 2. Item 7, not Items 2 and 3, appear to be the hardest, while Item 11 is the easiest. Unlike the tutors, only Items 2 and 3 fit the Rasch Model well. Of more interest is the lack of discrimination power of these items. All of the weighted fit MNSQ figures are less than the critical value of 0.77. This

means that these items only discriminate between students in a very narrow range of ability. Figure 3, below, shows that these items generally only discriminate between students in a very narrow range in the very low student ability range. Of particular concern is Item 9. It does not fit the Rasch Model well (unweighted fit T value of -3.80). This value suggests that the item is testing abilities or competencies that are markedly different to those that are being tested by the other 11 items. The same may also be said for Item 7, even though it does not exceed the critical value of -2.00 for this measure. Table 4 also shows that there is a little difference in the difficulty of the items. The range of the item estimates is only 0.292 logits. On the basis of this evidence there does not appear to be a significant difference in the difficulty of the items. Hence, the evidence in this regard does not tend to support student concerns about inter-item variability. Nevertheless, the specification if Items 7 and 9 need to be improved or the items should be deleted from the test.

Table 4: Items, Summary Statistics

VARIABLES UNWGHTED FIT WGHTED FIT

---

item DISCRIMINATION ERROR MNSQ T MNSQ T

---

1 0.051 0.029 0.62 -1.4 0.52 -8.1

2 -0.014 0.038 0.88 -0.2 0.73 -3.7

3 0.118 0.025 0.64 -1.5 0.61 -6.7

4 -0.071 0.034 0.75 -0.8 0.68 -4.9

5 -0.128 0.023 0.67 -1.5 0.54 -8.8

6 -0.035 0.034 0.84 -0.5 0.76 -3.7

7 0.148 0.025 0.58 -1.9 0.53 -10.5

8 0.091 0.030 0.72 -1.0 0.65 -5.9

9 0.115 0.019 0.39 -3.8 0.34 -17.5

10 0.011 0.032 0.74 -0.9 0.63 -5.4

11 -0.144 0.034 0.77 -0.8 0.66 -4.6

12 -0.142*

N=833

Figure 2: Map of Latent Distributions and Response Model Parameter Estimates.

Terms in the Model Statement

rater item rater by item

+**1** | | | |

| | | |

| | | |

| | | |

| | |8.5 4.8 4.9 |

| | |1.2 6.4 5.5 7.5 |

|2 3 5 |7 |2.1 2.2 6.2 1.3 |

|6 7 8 |1 3 8 9 10 |1.1 3.1 5.1 4.2 |

0 | |2 4 5 6 |4.1 6.1 7.1 8.1 |

|4 |11 12 |3.2 5.2 8.2 2.3 |

| | |7.2 6.3 1.4 8.4 |

| | |1.10 |

|1 | | |

| | |4.5 |

| | | |

-1 | | | |

N = 833

Vertical scale is in logits

Some parameters could not be fitted on the display

Figure 2 demonstrates some other interesting points. First, the closeness of the leniency of the majority of Raters and Items. The range in Item difficulty is only 0.292 logits. The most

interesting feature of this figure is that the maximum intra-rater variability. Rater 4, is approximately 50 per cent greater than the inter-rater variability. That is, the range of the inter-rater variability is 0.762 logits. Yet the intra-rater variability of Rater 4 is 1.173 logits as shown by the difference in the standard set for Item 5 (4.5 in Figure 2) and Items 8 and 9 (4.8 and 4.9 in Figure 2). This can be interpreted as the rater marking certain items at a level that would place a student in the top quartile, such as Item 5. But is then marking other Items of a similar difficulty, such as Items 8 and 9, at a level that would place the same student in the bottom quartile. It is interesting to note that the easiest marker, Rater 1, is almost as inconsistent as Rater 4, with an intra-rater variability of 0.848. With two notable exceptions, the intra-rater variation is less than the inter-rater variation. Nevertheless, intra-rater differences do appear to be significant. On the basis of this limited evidence it may be concluded that intra-rater variability is as much a concern as inter-rater variability. It also appears that intra-rater variability is directly related to the extent of the variation from the standard set by the subject coordinator.

Figure 3: Item Characteristic Curve, Item 2

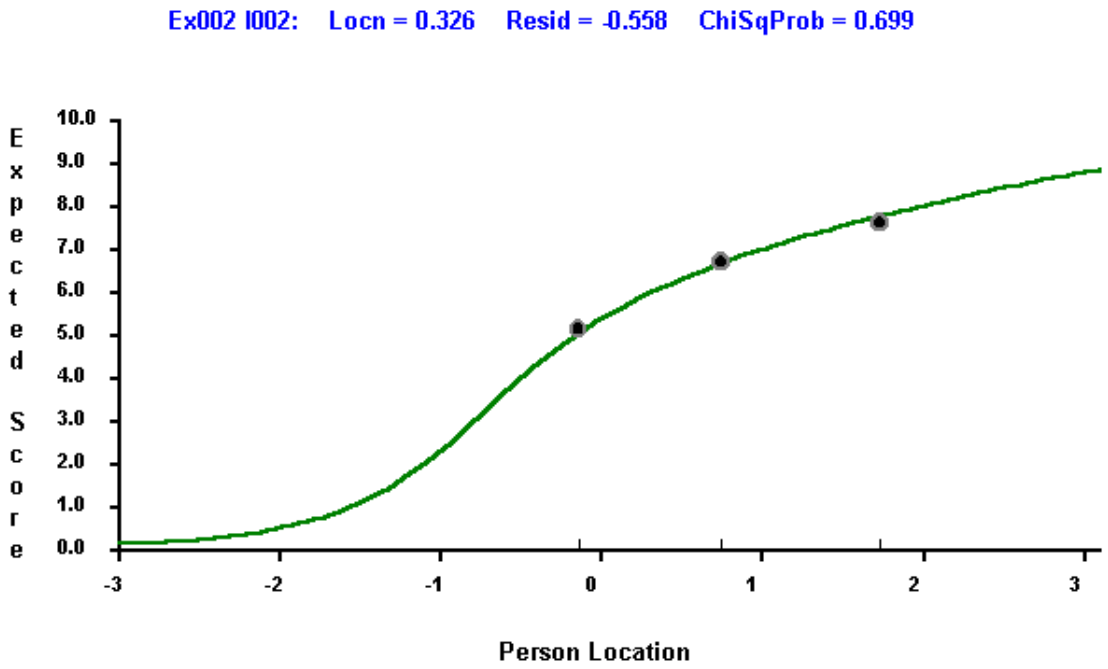
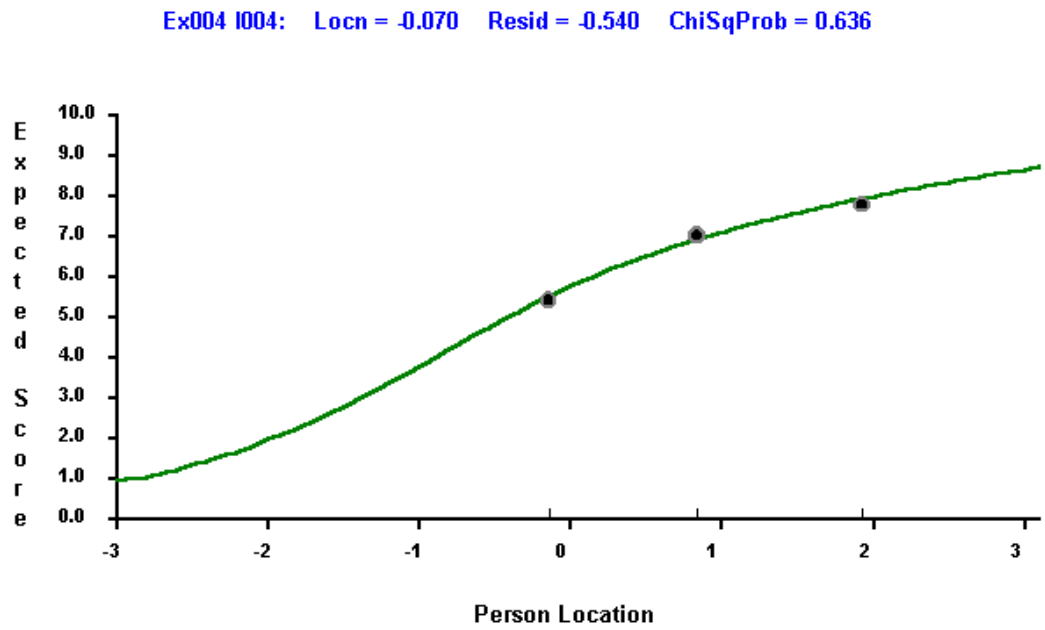


Figure 4: Item Characteristic Curve, Item 3



The Item Characteristics Curves that were obtained from RUMM confirm the item analyses that were obtained from Conquest. Figure 3 shows the Item Characteristic Curve for Item 2, which is representative of 11 of the 12 items on this exam. These items discriminate between students in a narrow range at the lower end of the student ability scale, as shown by the weighted fit MNSQ value being less than 0.77 for all items. However, none of these 11 curves has an expected value greater than much over 0.9. That is, the best students are not consistently getting full marks for their answers. This reflects the widely held view that inexperienced markers are unwilling to award full marks for essay questions. But on the other hand, Item 3 (Figure 4) discriminates poorly between students regardless of their ability. The weakest students are able to obtain quite a few marks, yet the best students are even less likely to get full marks that on the other 11 items. Either the item or its marking guide needs to be modified, or the item should be dropped from the paper. But more importantly, all of the items, or their marking guides, need to be modified in order to improve their discrimination power.

In short, there is little correspondence between the results obtained by analysing these data using Excel or the Rasch Model. Consequently, any actions taken to improve either the item or test specification, based on the Excel analysis, could have rather severe unintended consequences. However, the analysis needs to be repeated with some crossover between tutorial groups in order to separate any effects of the relationships between students and raters. For example, Rater 6 may only appear to be the toughest marker as his tutorials have an over representation of weaker students. These effects can be investigated by looking at the second set of results from the double marked papers.

Phase three of three study

The second phase of this study was designed to maximise crossover between raters and items, but there was no crossover between raters and students. The results obtained in relation to rater leniency and item difficulty may be influenced by the composition of tutorial groups as students are not be randomly allocated to tutorials. Hence, a 20 per cent sample of papers were double marked in order to achieve the required cross-over and provide some

insights into the effects of fully separating, raters, items and students. Results of this analysis are summarised in Tables 5 and 6 Figure 5.

Table 5: Raters, Summary Statistics

VARIABLES UNWGHTED FIT WGHTED FIT

---

rater LENIENCY ERROR MNSQ T MNSQ T

---

1 -0.123 0.038 0.92 -0.1 0.84 -0.8

2 0.270 0.035 0.87 -0.2 0.83 -1.0

3 -0.082 0.031 0.86 -0.3 0.82 -1.1

4 0.070 0.038 1.02 0.2 0.91 -0.4

5 -0.105 0.030 1.07 0.3 1.09 0.6

6 0.050 0.034 0.97 0.1 0.95 -0.2

7 0.005 0.032 1.06 0.3 1.04 0.3

8 -0.085*

N= 164

The first point that emerges from Table 5 is that the separation of raters, items and students leads to a reduction in inter-rater variability from 0.762 logits to 0.393 logits. Rater 1 is still the most lenient. More interestingly, Rater 2, the subject coordinator has become the hardest marker, reinforcing her status as the 'expert'. This separation has also increased the error for all tutors, yet at the same time reducing the variability between all eight raters. More importantly all eight raters now fit the Rasch Model quite well as shown by the unweighted fit statistics. In addition all raters are now in the critical range for the weighted fit statistics. Hence, they appear to be discriminating between students over an acceptable range of ability.

Table 6: Items, Summary Statistics

VARIABLES UNWGHTED FIT WGHTED FIT

---

item DISCRIMINATION ERROR MNSQ T MNSQ T

---

1 0.054 0.064 1.11 0.4 1.29 1.4

2 0.068 0.074 1.34 0.7 1.62 2.4

3 -0.369 0.042 0.91 -0.1 0.95 -0.3

4 0.974 0.072 1.33 0.7 1.68 2.8

5 -0.043 0.048 1.01 0.2 1.11 0.8

6 -0.089 0.062 1.10 0.4 1.23 1.2

7 -0.036 0.050 0.92 -0.1 1.02 0.2

8 -0.082 0.050 0.99 0.1 1.07 0.5

9 -0.146 0.037 0.75 -0.7 0.80 -1.5

10 0.037 0.059 1.01 0.2 1.13 0.7

11 -0.214 0.057 1.14 0.4 1.41 2.1

12 -0.154*

N = 164

However, unlike the rater estimates, the variation in item difficulty has increased from 0.292 to 1.43 logits (Table 6). Clearly now decisions about which questions to answer may be important determinants of student performance. For example, the decision to answer Item 4 in preference to Items 3, 9, 11 or 12 could see a student drop from the top to the bottom quartile, such is the observed differences in item difficulties. Again the separation of raters, items and students has increased the error term. That is, it has reduced the degree of consistency between the marks that were awarded and student ability. All items now fit the Rasch Model. The unweighted fit statistics, MNSQ and T, are now very close to one and zero respectively. Finally, ten of the weighted fit statistics now lie in the critical range for the weighted MNSQ statistics. Hence, there has been an increase in the discrimination power of these items. They are now discriminating between students over a much wider range of ability.

Finally, Figure 5 shows that the increased inter-item variability is associated with an increase in the intra-rater by item variability, despite the reduction in the inter-rater variability. The range of rater by item variability has risen to about 5 logits. More disturbingly, the variability for individual raters has risen to over two logits. Again Raters 1 and 4 are setting markedly different standards for the items that are of the same difficulty level. For example, Rater 1 is marking Item 4, the hardest item by far on this test, as if it was the easiest Item on the paper. The results obtained in this phase of the study differ markedly from the results obtained during the preceding phase of the study.

In general, raters and items seem to fitting the Rasch Model better as a result of the separation of the interactions between raters, items and students. But on the other hand, the intra-rater variability has increased enormously. However, the MNSQ and T statistics are a function of the number of students involved in the study. Hence, the reduction in the number of papers analysed in this phase of the study may account for much of changes in the fit of the Rasch Model with respect to the raters and items.

It may be concluded from this analysis that when students are not randomly assigned to tutorial groups then the clustering of students with similar characteristics in certain tutorial groups is reflected in the performance of the rater. However, a 20 per cent sample of double marked papers is too small to determine the exact nature of the interaction between raters, items and students.

Figure 5: Map of Latent Distributions and Response Model Parameter Estimates---------
---

Terms in the Model Statement

rater item rater by item

+4 | | | |

| | | |

| | | |

| | |2.4 |

| | | |

| | | |

| | | |

+3 | | | |

| | | |

| | | |

| | | |

| | | |

| | | |

| | | |

| | | |

+2 | | | |

| | | |

| | | |

| | | |

| | |4.4 |

| | |1.10 |

| | | |

+1 | |4 | |

| | | |

| | |5.3 1.6 |

| | |8.8 6.9 6.12 |

| | |3.1 1.2 7.6 6.7 |

|2 | |3.3 8.4 3.5 5.5 |

| | |7.1 4.2 6.3 8.3 |

|4 6 7 |1 2 10 |2.1 5.1 2.2 8.2 |

0 |1 3 5 8 |5 6 7 8 |1.1 4.1 6.2 7.3 |

| |9 11 12 |6.1 3.2 7.2 4.3 |

| |3 |5.2 2.5 4.5 7.7 |

| | |8.1 1.3 2.3 2.7 |

| | |5.4 2.6 |

| | |4.6 2.12 |

| | |2.8 4.10 |

-1 | | |7.4 |

| | |3.4 6.4 |

| | | |

| | | |

| | |1.4 |

| | | |

| | | |

| | | |

-2 | | | |

| | | |

N = 164

Conclusion

The literature on performance appraisal identifies five main types of rater error; severity or leniency, the halo effect, the central tendency effect, restriction of range and inter-rater reliability or agreement. This analysis has identified four of these types of errors applying to a greater or lesser extent to all raters, with the exception of the subject coordinator.

Clearly Rater 1, and to a lesser extent Rater 4, mark far more leniently than either the subject coordinator or the other raters. Second, there is however, no clear effect of the halo effect being present being present in the first Rasch Model (Table 2). Third, there is some evidence, Table 2 and Figures 2 and 3, the presence of the central tendency effect. Fourth, the weighted fit MNSQ statistics for the items (Table 3) shows that the items discriminate between students over a very narrow range of ability. This is also strong evidence for the presence of restriction of range. Finally, Table 2 provides evidence of unacceptably low levels of inter-rater reliability. Three of the eight raters exceed the critical value of 1.5, while a fourth is getting quite close.

These results demonstrate the need for some moderation in the marking of this examination to take into account the presence of these errors. However, a moderation process would conflict too much with the culture of the Division. Grades are awarded on raw not moderated scores. Despite the fact that staff are aware of the extent of moderation in certain Year 12 subjects. Nevertheless, these results can inform the staff development process and assist the University's staff developers to develop appropriate staff development courses. However, this may not be enough. It is necessary to ask questions about the subject coordinator, why is she so free of error and what can be done to make the other raters more like her? It may simply be the concept of ownership. She has a long-term employment relationship with the University, some status in the hierarchy and this, after all, is her subject. If increasing the level of ownership for raters is the key, then the solution to reducing inter-rater variability and intra-rater variability does not lie in the staff development process, but rather in the employment relations sphere of the University's activities.

Bibliography

Adams, R.J. and Khoo S-T. (1993) *Conquest: The Interactive Test Analysis System,* ACER Press, Canberra.

Andrich, D. (1978) A Rating Formulation for Ordered Response Categories, *Psychometrica*, 43, pp 561-573.

Andrich, D. (1985) An Elaboration of Guttman Scaling with Rasch Models for Measurement, in N. Brandon-Tuma (ed.) *Sociological Methodology*, Jossey-Bass, San Francisco.

Andrich, D. (1988) *Rasch Models for Measurement*, Sage, Beverly Hills.

Chase, C.L. (1978) *Measurement for Educational Evaluation*, Addison-Wesley, Reading.

Choppin, B. (1983) *A Fully Conditional Estimation Procedure for Rasch Model Parameters*, Centre for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.

Engelhard, G.Jr (1994) Examining Rater Error in the Assessment of Written Composition With a Many-Faceted Rasch Model, *Journal of Educational Measurement*, 31(2), pp 179-196.

Engelhard, G.Jr and Stone, G.E. (1998) Evaluating the Quality of Ratings Obtained From Standard-Setting Judges, *Educational and Psychological Measurement*, 58(2), pp 179-196.

Hambleton, R.K. (1989) Principles of Selected Applications of Item Response Theory, in R. Linn, (ed.) *Educational Measurement, 3rd ed.*, MacMillan, New York, pp. 147-200.

Keeves, J.P. and Alagumalai, S. (1999) New Approaches to Research, in G.N. Masters and J.P. Keeves, *Advances in Educational Measurement, Research and Assessment*, pp. 23-42, Pergamon, Amsterdam.

Rasch, G. (1968) A Mathematical Theory of Objectivity and its Consequence for Model Construction, European Meeting on Statistics, Econometrics and Management Science, Amsterdam.

Rasch, G. (1980) *Probabilistic Models for Some Intelligence and Attainment Tests*, University of Chicago Press, Chicago.

Saal, F.E., Downey, R.G. and Lahey, M.A (1980) Rating the Ratings: Assessing the Psychometric Quality of Rating Data, *Psychological Bulletin*, 88(2), 413-428.

van der Linden, W.J. and Eggen, T.J.H.M. (1986) An Empirical Bayesian approach to Item Banking, *Applied Psychological Measurement*, 10, pp. 345-354.

Sheridan, B., Andrich, D. and Luo, G. (1997) *RUMM User's Guide*, RUMM Laboratory, Perth.

Snyder, S. and Sheehan, R. (1992) The Rasch Measurement Model: An Introduction, *Journal of Early Intervention*, 16(1), pp. 87-95.

Weiss, D. (ed.) (1983) *New Horizons in Testing*, Academic Press, New York.

Weiss, D.J. and Yoes, M.E. (1991) Item Response Theory, in R.K. Hambleton and J.N. Zaal (eds) *Advances in Educational and Psychological Testing and Applications*, Kluwer, Boston, pp 69-96.

Wright, B.D. and Masters, G.N. (1982) *Rating Scale Analysis*, MESA Press, Chicago.

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