Vol.1 No.2 (2018)

Journal of Applied Learning & Teaching

Content Available at : **http://journals.sfu.ca/jalt/index.php/jalt/index**

Evaluating assessment in a School of Economics

Steven Barrett[A]
Francisco Ben[B]

A *School of Economics, The University of Adelaide*
B *School of Education, Tabor College of Higher Education*

Keywords

Assessment;
Inter-rater reliability;
Rubrics;
Performance criteria;
Partial credit model;
Sessional lecturers.

Article Info

Received 26 March 2018
Received in revised form 31 August 2018
Accepted 3 September 2018
Available online 14 December 2018

DOI: https://doi.org/10.37074/jalt.2018.1.2.2

Abstract

Previous research undertaken by one of the authors identified a general concern among undergraduate students in large business courses associated with the large number of sessional staff. In particular, students expressed their view in the focus groups with staff their concern that a large number of markers may affect their performance in essay style examinations as a result of the inevitable variation in severity between different raters.

The aim of this paper is to investigate whether these concerns are justified. The focus of this study was the weekly tutorial papers that were submitted for marking by 400 students enrolled in a first year Principles of Microeconomics course. These papers were marked by a team of ten markers whose experience in university teaching ranged from four weeks to over 30 years. 40 per cent of these papers were triple marked by two other raters in order to fully separate the student by rater by item interactions during the subsequent statistical analysis. The results obtained from this triple marking exercise were then analysed using ConQuest 2.0, which uses logistic regression to provide estimates of the parameters of the Partial Credit Model. The Partial Credit Model measures variations in rater severity and four other common rater errors, the halo effect, the central tendency effect, the restriction of range effect and the inter-rater variability or consistency.

The study identified the presence, to some degree or other, of all five rater errors, even among the most experienced raters. The paper concludes by suggesting that the key to improve rater performance lies in the design of marking rubrics.

Introduction

The School of Economics at the University of Adelaide runs a very large first-year principles programme mostly in response to the needs of service teaching in commerce programmes. Over the past decade, about 800 students enrolled in Microeconomics in Semester 1 and a further 300 students enrolled in the subject in Semester 2. On the other hand, about 300 students took Macroeconomics in Semester 1, and a further 800 students were enrolled in Macroeconomics in Semester 2. These two courses are taught using a conventional lecture/tutorial approach. In 2008, it was decided that in order to better align the teaching and learning activities of the School with the graduate attributes of the University, students were required to prepare a written answer to one of the weekly tutorial questions. In total students wrote ten tutorial papers and the best eight were counted towards their final assessment. These questions were marked by the tutors and the results accounted for ten per cent of the total mark for both courses.

Like in many other Australian universities, the School of Economics has responded to increased student enrolments and increasingly tight teaching budgets by increased flexibility in the employment relationships with its teaching staff. Most importantly, the School employs a large number of Honours and post-graduate research students as well as a small number of very experienced casual teaching staff as tutors. However, because of the constraints imposed by the University's Enterprise Bargaining Agreement at the time this research was undertaken, casual staff are only allowed to teach a maximum of five hours per week. So, large classes have a large team of tutors. In this case, a class of 800 students with 45 tutorial classes had 15 tutors.

Previous studies undertaken by the authors suggest that students are concerned when a course involves large numbers of markers (Barrett, 2005). In particular, students are concerned that a lack of consistency between markers may adversely affect their overall grade. However, the literature argues that marker consistency is just one of a number of rater errors that may affect student performance. The aim of this study was to analyse the results of the written tutorial answers for a one semester Microeconomics course, to identify the presence, or otherwise, of rater errors. The marks for these tutorial assignments were analysed using the Partial Credit Model (Andrich, 1978), which is a development of the Rasch Model (Rasch, 1968). Masters (1988) describes the Partial Credit Model as a latent trait or general polychotomous item response model that belongs to the Rasch family of latent trait models. In this study, logistic regression analysis is used to provide estimates of the parameters of the Partial Credit Model, that is rater severity, item difficulty and student ability. Moreover, the outputs of the Partial Credit Model can also be used to identify the presence of five common rater errors. This information can then be used to help raters avoid these errors in the future (Barrett, 2005).

In this project, the rating performance of the tutors involved in teaching this course was evaluated in order to identify the presence, or otherwise, of these five common rating errors as a basis for a course and staff development process. This project was very much a pilot study that was designed to explore these relationships.

This study analysed the rating performance of a sub-set of the tutors that helped deliver a first year Microeconomics course. Due to budget constraints, only 10 of the 15 tutors who were involved in teaching this course were invited to participate in this study. These tutors were chosen on the basis of two criteria. First, the authors were looking for a group of tutors that had the widest possible range of teaching experiences. Indeed, for one tutor (Rater 10) this was her first ever teaching experience, whereas the tutor-in-charge (Rater 2) is a very experienced teacher who has over 30 years of experience teaching Economics at both secondary school and university level. Secondly, in order to provide reliable estimates of the parameters of the Partial Credit Model, the study needed to analyse the ratings obtained from at least 350 students, preferably 400. So, the combination of tutors was chosen that would minimise the number of raters who taught 400 students.

The study explored the concerns that students have about marker consistency by using the Partial Credit Model to detect the presence (or otherwise) of five common rater errors. The following section is a brief review of the five common rater errors that are the focus of this study. The third section is the methods section. The fourth section is divided into three sub-sections and discusses further details of the study. The final section discusses the results of the study and presents the conclusions. This paper concludes that marking guides may not be sufficient to eliminate the five rater errors explored in this paper. Moreover, the paper suggests that these errors may be reduced by the use of marking rubrics. Hence, the paper concludes with a call for further research on this topic.

Five Rating 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, which is probably best understood as consistency (Saal, Downey & Lahey, 1980). Engelhard and Stone (1998) have demonstrated that the statistics obtained from the Partial Credit Model can be used to measure these five types of error. This section briefly outlines these rating errors 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 from the Partial Credit Model. The critical values reported in Table 1, relate to the rater and item estimates obtained from a statistical package called ConQuest 2.0 (Adams & Khoo, 1993), which is one of a number of commercially available software packages that can be used to analyse examination performance using either the Rasch Model or its extensions, such as the Partial Credit Model. The present study extends this procedure by demonstrating how Item Characteristic Curves and Person Characteristic Curves can also be used to identify these rating errors.

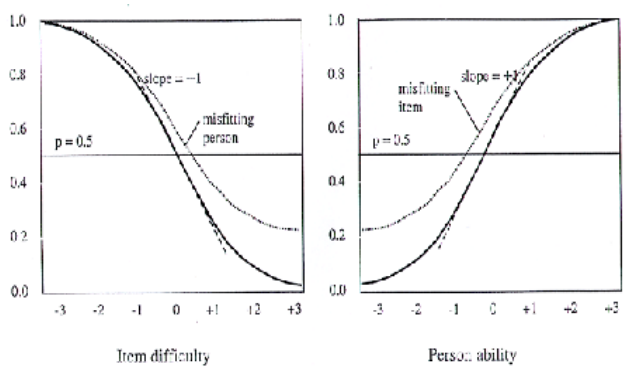

Figure 1: Item and Person Characteristic Curves
(Source: Keeves & Alagumalai, 1999, p.30).

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 that are produced by ConQuest 2.0 (Table 3 is an example of these statistics). The estimates for each rater should be compared with the expert in the field, or the standard setting judge. In this instance the tutor-in-charge, that is Rater 2, with over 30 years teaching experience, should be considered as the standard setting judge. If the leniency estimate of a particular rater is higher than the expert, then the rater is a harder marker, and if the estimate is lower, then the rater is an easier marker. Hence, the leniency estimates produced by ConQuest are reverse scored.

Evidence of rater severity or leniency can also be seen in the Person Characteristic Curves of the raters that are produced by software packages such as RUMM (Sheridan, Andrich & Luo, 1997). An example is provided in Figure 1. If the Person Characteristic Curve for a particular rater lies to the right of that of the expert 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 than the expert (Figure 1). Conversely, the differences in the difficulty of items can be determined from the estimates of discrimination produced by ConQuest. Table 4 provides examples of these estimates.

The halo effect

The halo effect appears when a rater fails to distinguish between conceptually distinct and independent aspects of student answers (Thorndike, 1920). For example, a rater may be rating items based on an overall impression of each answer, or be distracted by extraneous things such as handwriting. Hence, the rater may fail to distinguish between conceptually essential or non-essential material. The rater may also be unable to assess competence in the different domains or criteria that the items have been constructed to measure (Engelhard, 1994). Such a holistic approach to rating may also artificially create dependency between items. Hence, items, or parts of items in the case of multipart questions, may not be rated independently of each other. The lack of independence of rating between items can also be determined from the Partial Credit Model.

Evidence of a halo effect can be obtained from the Partial Credit Model by examining the rater estimates, in particular, the mean square error statistics, or weighted fit MNSQ. See Table 3 for an example. If these statistics are very low, that is less than 0.6, then raters may not be rating items independently of each other.

The shape of the Person Characteristic Curve for the raters can also be used to demonstrate the presence or absence of the halo effect. A flat curve, with a vertical intercept significantly greater than zero or which is tending towards a value significantly less than one as item difficulty rises, is an indication of the halo effect (Figure 1).

The central tendency effect

The central tendency effect describes situations in which the ratings are clustered around the mid-point of the rating scale and reflects reluctance by raters to use the extreme ends of the rating scale. This is particularly problematic when using a polychotomous rating scale, such as the one used in this study. The central tendency effect is often associated with inexperienced and less well-qualified raters.

This error can simply be detected by examining the marks of each rater using descriptive measures of central tendency such as the mean, median, range and standard deviation, but as illustrated in the fifth section of this paper, this can lead to errors. Evidence of the central tendency effect can also be obtained from the Partial Credit Model by examining the item estimates. In particular, the mean square error statistics, or unweighted fit MNSQ and the unweighted fit *t*. If these statistics are high, that is the unweighted fit MNSQ is greater than 1.5 and the unweighted fit *t* is greater than 1, then the central tendency effect is present. Central tendency can also be seen in the Item Characteristic Curves, especially if the highest ability students consistently fail to attain a score of one on the vertical axis and the vertical intercept is significantly greater than zero.

Restriction of range

The restriction of range effect is related to the central tendency effect as it reflects the reluctance of raters to use the extreme ends of the marking scale. It is also a measure of the extent to which the obtained ratings discriminate between different students with respect to their different performance levels (Engelhard, 1994; Engelhard & Stone, 1998). The underlying question that is addressed by restriction of range indices 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. Again, this error is associated with inexperienced and less well-qualified raters.

Evidence of the restriction of range effect can be obtained from the Partial Credit Model by examining the item estimates. In particular, the mean square error statistics, or weighted fit MNSQ. This rating error is present if the weighted fit MNSQ statistic for the item is greater than 1.30 or less than 0.77.

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 the Item Characteristic 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 Item Characteristic Curve will be very flat with little or no steep middle section to give it the characteristic "S" shape. Such an item fails to discriminate effectively between students of differing ability.

Inter-rater reliability or agreement

Inter-rater reliability or agreement, or consistency as it is more commonly known as, 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, this rating error reflects a concern with consensual or convergent validity. The model fit statistics obtained from the Partial Credit Model provides evidence of inter-rater reliability (Engelhard & Stone, 1998). It is unrealistic to expect perfect agreement with a group of raters. Nevertheless, it is not unrealistic to seek to obtain broadly 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 weighted fit MNSQ and 0.00 for the weighted fit *t* statistic. Weighted fit MNSQ greater than 1.5 suggest that raters are not rating items in the same order.

The unweighted fit 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.

Method

In the course that this study investigated, a total of 795 students were enrolled in 43 tutorials. Due to budgetary constraints, only a sub-set of 399 of these students from 28 tutorial groups and their ten tutors participated in the study. The tutors represented the full spectrum of experience. The tutor-in-charge is a retired secondary school economics teacher with some 10 years university teaching experience and 20 years secondary school teaching experience (Rater 2), another was a qualified high school teacher with more than 20 years university teaching experience (Rater 1), while for Rater 10 this was the first time she had ever taught at university. It was hoped that about 400 students would be involved in the study as 350 students is the minimum number of cases that are required by ConQuest to generate reliable estimates of the parameters of the Partial Credit Model. In particular, the *t* statistics are sensitive to sample size and require a sample size of at least 350. In the end, over 2,500 ratings from 399 students were analysed in this study.

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 that of the experts.	**Rater estimates.** Compare estimate of leniency with the expert. Lower error term implies more consistency.
Halo effect	Person Characteristic Curve. Maximum values do not approach 1 as student ability rises. Vertical intercept 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. Maximum values do not approach 1 as student ability rises.	**Item estimates;** Unweighted fit MNSQ >> 1 and 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 MNSQ <0.77 or Weighted fit MNSQ > 1.30.
Reliability	Person Characteristic Curve. Slope at point of inflection significantly greater than or less than 1.00.	**Rater estimates.** Weighted fit MNSQ >> 1 and Weighted fit *t* >> 0.

Table 1: Summary Table of Rater Errors and Rasch Test Model Statistics

At the heart of the Partial Credit Model is the premise that the performance of a student in essays is the interaction of student ability, the questions or items the student decides to answer and the markers. Hence, the Partial Credit Model uses logistic regression analysis to estimate the three parameters of the model, rater severity, item difficulty and student ability. A priori it would be expected that higher ability students should perform better than students of lower ability. However, if lower ability students choose to answer easier questions, or if more lenient raters mark their answers, then they may outperform more able students. This is the basis of student concerns. So, the aim of the Partial Credit Model is to separate the interactions between student ability, item difficulty and rater severity in order to properly evaluate student performance and rater performance. This separation between students, items and raters can only be achieved if there is crossover between students, items and raters. Crossover occurs when raters mark a range of questions and if they mark the work of students who are in tutorial groups other than their own in addition to their own students.

The tutors were given a briefing session about the project that lasted about an hour. The key part of the briefing was discussing the concept of crossover and why it was so important to this study. All of the tutors marked all of the tutorial questions for all of their students over the course of the semester. Students were required to submit written answers to ten tutorial questions, that is one question each week for Weeks two to eleven and the best eight were counted towards the final grade. Raters did not mark the work of students who were enrolled in other tutorials, so the crossover between raters and items was maximised. But, there was no crossover between raters and students. Thus, the study needed to develop a strategy such that

tutors would mark the papers submitted by students that belonged to tutorial groups other than their own to provide the crossover between raters, items and students that is required to obtain reliable estimates of the model parameters.

The standard response to obtaining crossover between raters and students is for a sample of about 20 per cent of the papers to be double marked by other members of the teaching team (Barrett, 2005). However, Englehard (1994) argued that this approach provides imprecise estimates of the parameters of the Partial Credit Model because some of the statistics are dependent on the sample size. Therefore, in order to produce reliable estimates of the model's parameters around 40 per cent of the papers were triple marked.

All of the tutorial papers that were marked by the ten tutors who took part in this study were photocopied twice prior to marking and around 40 per cent of these papers were allocated to two other tutors for double and triple marking. In the end, the marks for about 1,400 tutorial questions for 399 students were obtained. This is the first round of marking. Then a further 1,100 papers were double marked by a second rater and then triple marked by a third rater. The triple marking process was managed such that no person marked the same paper twice or indeed thrice. Clean papers were provided to the second and third markers so that they had no idea of the marks obtained from the first round of marking. Only the four tutorial questions that were submitted for marking in the second half of the semester were analysed. The tutors were paid to do extra marking. They were paid the relevant marking rate as per the University of Adelaide's Enterprise Bargaining Agreement. The marking was paid for out of a small grant from the School of Economics.

Training was undertaken in two phases. Prior to the initial round of marking, there was a meeting in which the marking guide was given to the tutors and then discussed. Then there was a second meeting prior to the second/third marking. This session was designed to inform the tutors about the project and to help them understand that the analysis was trying to unpack the item, rater-student interactions. There were no ethical issues that require discussion. The project was approved by the University of Adelaide's Human Research Ethics Committee.

The Study

This study investigated the concerns of students about essay marking. In particular, in large classes with large numbers of raters, variation in rater performance can adversely affect student outcomes. In order to ascertain the veracity of students' concerns, this study analysed four tutorial papers, submitted by 395 students and marked by ten raters. The analysis proceeded in three distinct phases. Phase 1 was an examination of the student mark using measures of central tendency. Such an approach is the norm to assess rater/marker performance. Phase 2 analysed the original marks of these 395 students using the Partial Credit Model. Such an approach may provide useful information. But the results will only be indicative due to the crossover effects. The rater by item crossover was maximised as the tutors marked all of the questions submitted by 'their' students. However, there is rater by student crossover. Phase 3 of the study maximised the rater by item crossover and the rater by student crossover by triple marking around 40 per cent of the papers. These results were then analysed using the Partial Credit Model.

Phase one of the study

The evaluation of assessment procedures at most Australian universities tends not to be very sophisticated. Indeed, it is rare "for researchers to consider the complex causal antecedents for observed rater effects" (Wolfe & McVay, 2012, p. 31). Typically, if the lecturer in charge of a large course with a large number of raters became concerned about rater consistency then the evaluation would be rather cursory. The lecturer would probably examine a range of measures of central tendency, such as the mean and the standard deviation. If these measures varied too much then the lecturer might be required to undertake some remedial action, such as moderation, staff development or even termination of those raters whose performance differed too much from the mean.

However, in subjects where students are free to enrol in whatever tutorial suits them, people with similar characteristics tend to be attracted to the same class. So, a tutor may be the only one taking the "after hours" classes, which may be attractive to older, more experienced students, that is, students with higher ability. They may also be more strategic learners and so have more successful strategies for addressing assessment activities. It should therefore not be surprising that this particular tutor's students also perform better. But such a tutor may be labelled "too lenient", requiring remedial action. Remedial action may have severe implications. It is time consuming for subject conveners and sessional staff may lose their jobs for no real reason. Hence, assessment evaluation needs to be undertaken properly and professionally to ensure that a problem exists in the first place.

	Question 1		Question 2		Question 3		Question 4		Average of
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	all items
Rater 1	4.77	1.71	6.00	1.81	7.72	1.93	6.67	2.54	6.25
Rater 2	8.31	1.89	5.95	1.43	9.00	1.06	7.06	1.51	7.60
Rater 3	7.91	2.92	5.73	2.40	7.12	3.72	7.82	2.11	6.95
Rater 4	7.22	1.53	4.38	1.60	7.00	2.55	8.81	1.97	6.80
Rater 5	7.20	1.20	7.08	1.22	8.68	1.80	7.18	1.33	7.50
Rater 6	8.74	1.08	7.57	1.67	9.07	2.13	8.25	1.43	8.45
Rater 7	6.62	2.06	5.40	2.06	7.74	2.28	6.33	1.86	6.50
Rater 8	7.06	1.03	6.63	1.03	8.16	1.19	7.4	1.28	7.30
Rater 9	7.19	1.10	6.69	1.12	8.50	1.43	6.73	0.88	7.20
Rater 10	8.00	1.38	8.0	1.38	7.71	2.65	8.69	1.23	8.00

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

The first phase of the present study analysed the mean marks and the standard deviation of the ten raters for the four items that they marked. The data presented in Table 2 reveals some interesting differences between raters that the tutor in charge of this course may want to consider.

An examination of Table 2 suggests that Rater 1 is a particularly hard marker with an average score of only 6.25

out of ten. Whereas Rater 2, the tutor-in-charge, seems to be about the middle of the range, as would be expected from the standard setting judge and the most experienced rater.

On the other hand, Rater 10 appears to one of the most lenient raters, which would normally not be unexpected given that this is the first time she had ever taught. This table shows that Rater 6 is the most lenient rater. Whereas, Raters 3 and 4 are providing ratings that are broadly consistent with the standard setting judge, which would be interpreted as being reliable raters. However, does this table draw the correct conclusions about the performance of these raters? The answer to this question is the focus of the next section.

Phase two of the study

The second phase of the study was a re-examination of the students' marks using the Partial Credit Model. It was noted above that the crossover between raters and items was maximised as all raters marked all items. However, there was no crossover between raters and students. Raters only marked the work of their students. In this phase of the study it was decided not to do anything to correct for this lack of crossover. Rather, it was decided to analyse the test results with no crossover between raters and students and compare them with the results obtained when the rater, item and student interactions are completely separated.

The information presented in Table 2 was re-analysed using the Partial Credit Model and is presented in Tables 3 and 4. The estimate for rater severity or leniency is presented in the first column of Table 3, while the estimates for item difficulty are provided in the first column of Table 4. The information presented in Tables 3 and 4 reflect the results of the triple marking process. Hence, there is maximum crossover between raters and items, but the influence of student ability on the measured severity of raters and the difficulty of the items is not included in these tables. The results from the Partial Credit Model are not very dissimilar to those presented in Table 2. This is not surprising, given the analysis that produced the information in Table 3 does not include the effects of the interaction of student ability on rater severity or item difficulty. The estimates for rater performance, item difficulty and student performance in all the following tables is derived from logistic regression analysis. The parameters are all measured in units called *Logits*. In a *Logit* scale, using the context of this study, "0" indicates average. Values above the average indicates markers tending to be harsh or that items are too difficult (more challenging). Whereas values below the average indicates lenient markers or easy items (less challenging).

			UNWEIGHTED FIT			WEIGHTED FIT		
Rater	ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
1	0.850	0.111	1.10	(0.26, 1.74)	0.4	1.07	(0.23, 1.77)	0.1
2	-0.664	0.107	0.75	(0.35, 1.65)	-0.7	0.88	(0.28, 1.72)	-0.4
3	-0.092	0.107	0.93	(0.33, 1.67)	-0.1	0.92	(0.29, 1.71)	-0.1
4	0.628	0.108	1.44	(0.33, 1.67)	1.2	1.43	(0.29, 1.71)	1.0
5	-0.976	0.108	0.90	(0.33, 1.67)	-0.2	0.92	(0.29, 1.71)	-0.2
6	0.518	0.103	1.02	(0.35, 1.65)	0.2	1.05	(0.31, 1.69)	0.1
7	0.022	0.105	0.81	(0.41, 1.59)	-0.6	0.76	(0.33, 1.67)	-0.7
8	0.207	0.103	1.04	(0.38, 1.62)	0.2	1.02	(0.34, 1.66)	-0.0
9	0.460	0.322	0.78	(0.35, 1.65)	-0.6	0.78	(0.31, 1.69)	-0.6
10	-0.954	0.112	1.08	(0.28, 1.72)	0.3	1.16	(0.21, 1.79)	0.3

Table 3: Rater Estimates Obtained from the Partial Credit Model

Rater 2, the standard-setting-judge has an estimate of severity of -0.664, an estimate of zero would be ideal. So, Rater 2 might be considered as somewhat lenient. Nevertheless, he fits the model quite well, as shown by his weighted and unweighted fit statistics. Both of his MNSQ estimates fall with the critical values and both t statistics are close to zero. Rater 10, with a leniency estimate of -0.954 is one of the most lenient raters, as might be expected. Rater 1 is still the most severe rater. On the other hand, Table 4 shows a substantial variation in item difficulty, over two *Logits*, which may have an effect on student performance in a situation where students are essentially free to choose which questions they answer. Nevertheless, all the items fit the model quite well.

			UNWEIGHTED FIT			WEIGHTED FIT		
Item	ESTIMATE	ERROR	MNSQ	CI	T	MNSQ	CI	T
WK7	-0.907	0.071	1.24	(0.73, 1.27)	1.7	1.31	(0.72, 1.28)	2.0
WK8	1.215	0.064	0.88	(0.78, 1.22)	-1.1	0.84	(0.78, 1.22)	-1.4
WK9	-0.784	0.066	1.11	(0.77, 1.23)	1.0	1.18	(0.77, 1.23)	1.4
WK11	0.475	0.117	0.83	(0.78, 1.22)	-1.5	0.81	(0.78, 1.22)	-1.8

Table 4: Item Estimates Obtained from the Partial Credit Model

As alluded to above, an important point needs to be made about Table 3. It provides estimates of rater severity taking into account only the inter-relationship between the rater and the items. Hence, it ignores the effect of student ability on the obtained ratings. Hence, Rater 1 may well be the hardest rater, but he may just have appeared to be the hardest rater in Tables 2 and 3 as his students decided to choose the most difficult questions. Or this rater had a higher proportion of lower ability students. In the context of this study, those students who find the linguistic challenges of economics more difficult than other students, might be over-represented in his tutorial groups. The inference that differences in rater performance might be conditional on the items answered or the composition of tutorial groups means that there is a need to fully separate the item, rater and student interactions. This led to the third phase of the study.

Stage three of the study

In this phase of the study, the item, rater and student interactions are completely isolated by maximising the crossover between raters and students. The crossover between raters and items was already maximised in both phases one and two of the study. The usual practice to obtain crossover in studies such as this is achieved by double marking around 20 per cent of papers. This is possible as the Partial Credit Model can still develop estimates of the parameters with missing data. This study was overly conservative. Hence, around 40 per cent of the papers were blind triple marked. Many of the estimates of the statistics obtained from the model are sensitive to sample size. So, the decision to triple mark 40 per cent of the papers meant that there was an eight fold increase in the number of "double marked" papers, which should commensurately increase the accuracy of the estimates of the statistics.

The focus of this section of the study is the information presented in Table 5. This information is derived from the triple marking. The triple marking allows for the interactions between item difficulty, rater severity and student ability to be separated. Table 5 is a map of student ability, rater

severity and item difficulty. These are the three parameters of the Partial Credit Model and are mapped onto the same scale using *Logits* as the unit of measurement. This Table separates the student, rater and item interactions and hence provides more accurate insights into rater severity and item difficulty. Moreover, the final column provides information about the interaction between the raters and the individual items. Hence, this column provides information about inter-rater variability, or consistency. Table 5 also shows a number of interesting points.

First, the tutor-in-charge (Rater 2) emerges as the hardest rater, with a severity estimate of about 1 *Logit*, which is in stark contrast to the estimate provided in Table 3 (-0.664). However, the average ability of the students in this course was about 4.25 *Logits*. So, even though he is the most severe rater, his ratings are comparatively easy compared to student ability.

Second, Rater 10, who had never taught before, now emerges as one of the more severe raters, marking about as severe as the standard setting judge. Hence, she is not the most lenient rater as suggested by Tables 2 and 3. Her apparent leniency could be explained in terms of the interaction of the ability of the students in her classes (higher) and the questions they chose to answer (easier).

Third, Rater 6 was shown as a lenient marker in Table 2 and a severe maker in Table 3. However, Table 5, taking into account all the student, rater and items interactions, shows that Rater 6 is indeed one of the more lenient raters. Whereas, Rater 1, who has consistently been shown to be the most severe rater now emerges as one of the more lenient raters. This might be explained in terms of the ability of the students in his classes and the items they answered. So, on balance it seems that these students tended to be lower ability, that is, they may have found the linguistic challenges of economics more challenging or they did not have well-developed study skills or they left their hand-up assignments until the last minute and so 'chose' to answer the more difficult items.

Finally, the estimates for item difficulty do not change much between Tables 4 and 5 as a result of fully taking into account the student, rater item interactions. The rater by item estimates show the extent of inter-rater variability, or consistency. The range of item difficulty shown in Table 4 is about 2.1 *Logits*, whereas the range of item difficulty shown in Table 5 is about 2.25 *Logits*. So, the full separation of students, raters and items does not affect the estimates for item difficulty too much. But what is more interesting is the range of item difficulty shown in the rater by item column. This column shows that the range in item difficulty is now about 4.25 *Logits*, which is double the range of item difficulty shown in the item column. This means that raters are marking items as if they are harder or easier than they really are.

For example, Rater 7 marked item 1 (point 7.1) as the most difficult item, when in fact it is the easiest item. Moreover, he marked it as if it was substantially harder than the actually hardest question, which is item 2. On the other hand, look at point 2.2. This shows that Rater 2 marked item 2, which is the most difficult item, as if it were considerably easier than the easiest item. Yet his other ratings were broadly consistent with the item difficulty. This lack of consistency is probably best explained in terms of the marking guides that were used by the raters. It would appear that they need to be re-designed to provide raters with more information about difference in item difficulty.

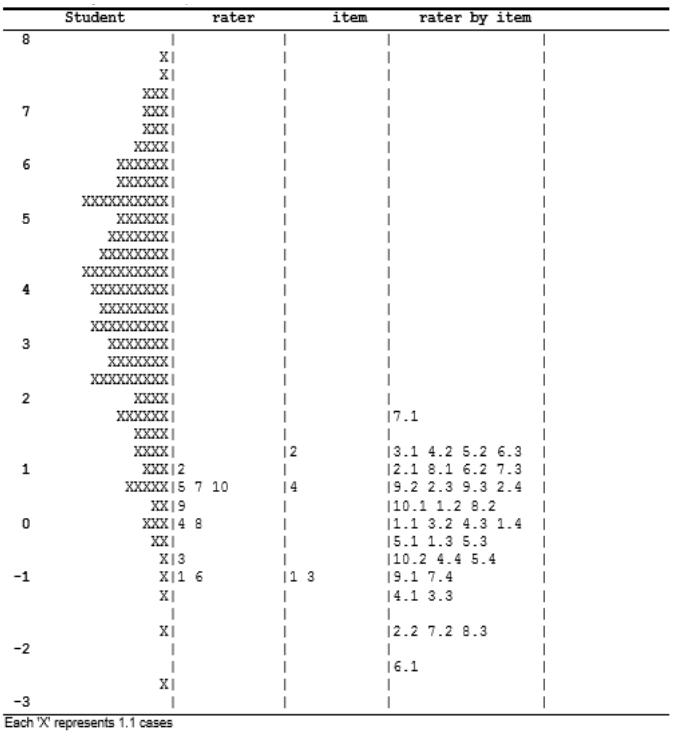

Table 5: Map of Student, Rater and Item Interactions

Table 3 can also be used to establish the presence of the halo effect. As shown in Table 1, the halo effect is present if the weighted fit MNSQ is less than one. Table 3 shows that four raters exhibit the halo effect to some extent, these are Raters 2, 3, 5, 7 and 9. Interestingly, Rater 10, the least experienced rater, has the second highest Weighted Fit MNSQ and hence does not exhibit the halo effect. Such a high incidence of the halo effect can probably be explained in terms of the marking guides. These results suggest that the marking guides at least need to be redesigned if not replaced by marking rubrics.

Table 1 shows that the central tendency effect is present if the unweighted fit MNSQ is greater than one and if the unweighted t is very much greater than zero. Table 6 indicates that about one third of the ratings were affected by the central tendency effect. Raters 1, 2, 9 and 10 were free from this error. It is not surprising that Raters 1 and 2 were free from this error, given their experience. It was a surprise to see that Rater 10 was also free from this error. On the other hand, Rater 4 exhibited the central tendency effect for three items, that is, questions 1, 2 and 3. Again the prevalence of this error may be reduced by the development of marking rubrics.

Table 2 suggests that raters whose weighted fit statistics fall outside of the critical interval demonstrate the restriction of range effect. Table 6 shows that the restriction of range effect occurs in 24 of the 40 rater by item statistics shown in Table 6. Again this finding suggests that the marking guides need further development.

Table 6 can be used to develop the above discussion about

rater	item	ESTIMATE	ERROR	UNWEIGHTED FIT MNSQ	UNWEIGHTED FIT T	WEIGHTED FIT MNSQ	WEIGHTED FIT T
1	WK7	0.988	0.162	0.70	-0.5	0.71	-0.6
2	WK7	1.116	0.151	0.68	-0.9	0.67	-1.0
3	WK7	-1.132	0.159	0.91	-0.1	0.69	-0.7
4	WK7	-0.292	0.157	1.39	0.9	1.41	0.9
5	WK7	-2.380	0.164	0.51	-1.1	0.58	-0.9
6	WK7	1.809	0.145	1.14	0.5	1.15	0.4
7	WK7	0.779	0.155	2.19	2.2	2.28	2.3
8	WK7	-0.965	0.155	3.63	3.6	3.84	3.8
9	WK7	0.145	0.471	0.50	-1.1	0.49	-1.2
10	WK7	-0.067	0.163	0.86	-0.1	0.85	-0.3
1	WK8	-1.717	0.147	0.84	-0.3	0.84	-0.3
2	WK8	-0.100	0.142	0.48	-1.7	0.47	-1.8
3	WK8	1.291	0.139	1.06	0.3	1.03	0.1
4	WK8	1.166	0.142	1.64	1.6	1.31	0.9
5	WK8	0.788	0.141	0.62	-1.1	0.64	-1.1
6	WK8	-1.797	0.142	1.10	0.4	1.09	0.3
7	WK8	0.297	0.135	0.99	0.1	0.99	0.0
8	WK8	0.342	0.134	1.00	0.1	1.02	0.1
9	WK8	-0.580	0.422	0.45	-1.7	0.44	-1.8
10	WK8	0.310	0.146	0.49	-1.5	0.50	-1.5
1	WK9	0.473	0.144	1.13	0.4	1.08	0.3
2	WK9	-1.186	0.148	0.49	-1.8	0.60	-1.1
3	WK9	0.376	0.144	1.01	0.1	1.04	0.2
4	WK9	-0.260	0.142	1.13	0.5	1.22	0.7
5	WK9	1.291	0.140	0.75	-0.7	0.71	-0.9
6	WK9	0.960	0.135	0.75	-0.6	0.77	-0.8
7	WK9	-1.916	0.149	0.53	-1.1	0.54	-1.2
8	WK9	0.473	0.139	4.27	5.2	4.19	5.2
9	WK9	0.007	0.430	0.34	-2.5	0.34	-2.5
10	WK9	-0.218	0.148	0.75	-0.6	0.86	-0.3
1	WK11	0.256	0.262	1.38	1.0	1.37	1.0
2	WK11	0.171	0.255	0.39	-2.0	0.40	-2.1
3	WK11	-0.535	0.255	1.11	0.4	1.11	0.4
4	WK11	-0.614	0.255	0.83	-0.4	0.73	-0.9
5	WK11	0.301	0.257	0.56	-1.3	0.54	-1.4
6	WK11	-0.972	0.243	0.48	-1.8	0.47	-1.9
7	WK11	0.840	0.254	1.73	1.7	1.67	1.5
8	WK11	0.150	0.248	0.77	-0.6	0.75	-0.7
9	WK11	0.429	0.765	0.58	-1.4	0.57	-1.5
10	WK11	-0.025	0.264	0.69	-0.8	0.69	-0.8

Table 6: Rater by Item estimates

inter-rater reliability of raters. As discussed in Table 1, the reliability error is present if the weighted fit MNSQ is greater than one and if the Weighted fit *t* is greater than zero. This error is evident in the performance of Raters 4, 6 and 10. However, this finding does not support the evidence that is provided in Table 5, which suggests that Rater 10 is quite reliable. The discrepancy between the results presented in Tables 5 and 6 can be explained by the fact that Table 5 shows the full student, by rater by item interactions. Again, rater reliability might be improved by developing the marking rubrics.

Concluding Remarks

This paper shows that evaluating rater performance is a much more difficult process than most academic managers expect. Proper rater evaluation needs to be more sophisticated than a cursory examination of measures of central tendency. Moreover, it is surprisingly easy to make an incorrect evaluation of rater performance as most managers would not be able to separate the complex interactions between student ability, the difficulty of questions and rater performance. This paper used the Partial Credit Model to evaluate one particular aspect of rater performance, the presence, or otherwise, of five common rater errors among a team of ten tutors. The team that was investigated here was a rather diverse group of people with varying levels of experience teaching First Year university Economics courses, ranging from just a few weeks to 30 years.

It comes as no surprise to the authors that the five common rater errors were present in the ratings of all ten raters. However, it appears that the more experienced raters were less prone to making these errors. But make these errors they did. The surprising finding of this study is that the least experienced rater, the rater for whom this was her first ever teaching job, was relatively free from making these errors. The new tutor was an exceptional young woman. She was a German international student doing Honours with the University at the time the study was conducted. It was planned to work out why she was such an effective marker, which might inform the tutor training process. Unfortunately, by the time the results of the project had become available she had been offered a PhD scholarship at another interstate university and lost contact. Further exploration of the explanation of the study's key findings would entail a derivative study.

Previous work by the authors suggested that the presence, or otherwise of these five errors was related to the nature of the employment relationships of the rater and the concept of ownership. That is, raters who were tenured or were employed on a long-term contract tend to have more ownership of courses and hence were less prone to making these particular five errors. In this study all ten raters were employed on a casual/sessional basis. So, the concept of ownership may not be appropriate. Nevertheless, this study would suggest that large classes should be taught by as few tutors as possible, teaching as many classes as practicable. However, this is not always possible. Although these raters were provided with comprehensive marking guides, they were not provided with marking rubrics. Even the most comprehensive marking guide still provides raters with some discretion or latitude, which in turn may create the space for rater errors to emerge. It would therefore be interesting to replicate this project to investigate whether the use of marking rubrics reduced the frequency and extent of these five rating errors.

The key finding of the study is that there appears to be a rating gradient. People who have been teaching and marking longer tend to make fewer rating errors than people with less experience. However, given the dynamics of the tutor / marker workforce, which turns over very quickly, most tutors do not get the experience to be relatively error-free. So, the challenge is to help people who are going to be tutors for a few years while they do their PhDs reduce their propensity to make errors. The key to this seems to be the development of better / clearer marking criteria or rubrics to support inexperienced tutors. Another suggestion would be to provide tutors with some form of professional development (a more in-depth training) in marking assessment tasks.

As a final word, we would like to re-visit Figure 1. The underlying premise of the Partial Credit Model is that student performance on essay style examinations is the outcome of the interaction between students, items and raters. The output of the Partial Credit Model includes the Item Characteristic Curve and the Person Characteristic Curve. These curves allow the performance of items and students to be compared to the model in order to identify misfitting items and students. However, at present there is no simple way to identify misfitting raters. So, this paper concludes with a call to the authors of the software to develop a Rater Characteristic curve in order to identify misfitting raters.

References

Adams, R. J. & Khoo, S-T. (1993). *Conquest: The interactive test analysis system*. Canberra, Australia: ACER Press.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrica, 43*, 561-573.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement, in N. Brandon-Tuma (Ed.) *Sociological Methodology.* San Francisco, CA: Jossey-Bass.

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills, CA: Sage.

Barrett, S. R. F. (2001). The impact of training in rater variability. *International Education Journal 2*(1), 49-58.

Barrett, S. R. F. (2001). Differential item functioning: A case study from first year economics. *International Education Journal, 2*(3), 1-10.

Barrett, S. R. F. (2005). Raters and examinations. In S. Alagumalai, D. A. Curtis & N. Hungi (Eds.), *Applied research measurement: A book of exemplars*: *Papers in honour of John P. Keeves* (pp. 159-177). New York, NY: Springer.

Chase, C. L. (1978). *Measurement for educational evaluation*, Reading, England: Addison-Wesley.

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), 179-196.

Engelhard, G. Jr., & Stone, G. E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement, 58*(2), 179-196.

Hambleton, R. K. (1989). Principles of selected applications of item response theory. In R. Linn, (Ed.), E*ducational measurement* (3rd ed.), (pp. 147-200). New York, NY: MacMillan.

Keeves, J. P., & Alagumalai, S. (1999). New approaches to research. In G. N. Masters & J. P. Keeves, *Advances in educational measurement, research and assessment* (pp. 23-42). Amsterdam, The Netherlands: Pergamon.

Masters, G. N. (1988). Partial credit models. In J. P. Keeves (Ed.), *Educational research methodology, measurement and evaluation* (pp. 292-296). Oxford, England: Pergamon Press.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory* (pp. 101-121). New York, NY: Springer.

Rasch, G. (1968). A mathematical theory of objectivity and its consequence for model construction. Copenhagen, Denmark: *European Meeting on Statistics, Econometrics and Management Science*.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

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.

Sheridan, B., Andrich, D., & Luo, G. (1997). *RUMM user's guide*, RUMM Laboratory, Perth.

Snyder, S., & Sheehan, R. (1992). The Rasch measurement model: An introduction. *Journal of Early Intervention, 16*(1), 87-95.

van der Linden, W. J., & Eggen, T. J. H. M. (1986). An empirical Bayesian approach to item banking. *Applied Psychological Measurement*, *10*, 345-354.

Weiss, D. J. (Ed.). (1983). *New horizons in testing*. New York, NY: Academic Press.

Weiss, D. J. & Yoes, M. E. (1991). Item response theory. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in Educational and psychological testing and applications* (pp. 69-96). Boston, MA: Kluwer.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantially interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31-37.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., & Stone M. H. (1979). *Best test design*. Chicago, IL: MESA Press.