Psychological Test and Assessment Modeling, Volume 59, 2017 (4), 453-470

Some IRT-based analyses for interpreting rater effects

Margaret Wu[1]

Abstract

In this paper, we present a few IRT-based analyses of rater effects including an examination of rater severity and rater discrimination. Rater severity refers to the differences between raters in terms of their tendencies to award higher or lower scores. Rater discrimination refers to the extent to which raters use the score range to separate students on the ability scale. Methodologies to estimate rater severity and rater discrimination are presented. A discussion on the interpretations of some measures of rater effect is provided. We highlight that a rater who shows large discrepancies from other raters may in fact be the best rater.

Keywords: rater severity, central tendency, rater discrimination

[1] *Correspondence concerning this article should be addressed to:*Margaret Wu, PhD, Assessment Research Centre, Melbourne Graduate School of Education, The University of Melbourne, Victoria 3010, Australia; email: wu@edmeasurement.com.au

Introduction

Many assessment types require raters to make judgements of the quality of the tasks students performed. These assessment tasks may include essay writing, music performance, artwork production, presentations and many other task forms that require raters' holistic judgements based on a wide range of considerations. In this paper, we will use essay marking as the basis for our discussion of rater effects. Of course, the analyses described below can be applied equally to other performance tasks.

First we note that assigning scores to performance tasks is a subjective process. Raters typically need to take into account many different aspects of a student's work in assigning a score. Even with rater training and monitoring programs in place, raters inevitably bring their own perspectives, emphases, interpretations and experiences to the judging process given that every piece of students' work is likely to be original and unique (e.g., Cook et al., 2009; Barret, 2001; Weigle, 1998). For this reason, we will refrain from using the term "rater error" which suggests that there are correct ratings and incorrect ratings. Rather, we will use "rater effects" or "rater characteristics" to describe different rater behaviours. We will, however, explain that some rater effects are more important than others, in terms of making the assessments more "useful" such as providing more information for the stakeholders of the assessments. In some assessments of essays, a group of experts (reference raters) have rated a number of essays and the experts' ratings are regarded as the "correct" scores. Scores of other raters are compared with those of the experts. Based on our experience and others (e.g., Attali, 2016; Leckie & Baird, 2011), we note that expert raters will also bring their own perspectives, and different experts typically will not always agree with each other. Therefore, the selection of a group of experts can also be a somewhat arbitrary decision in the assessment process.

In this paper, we will not use a group of "reference raters" as the basis of comparison for other raters. We will compare a rater's scores with those of the rest of the raters as a group, thus demonstrating relative rater effects. For example, if a rater is harsh, it means that he/she provides lower ratings as compared to other raters in the group. It does not necessarily mean that a harsh rater is one who made errors in his ratings. We will explain this point further later in the paper.

To clarify the terminology used in this paper, we will use the term "items" to refer to either different essays a student has written or different criteria a rater has to provide scores for based on one essay. That is, if there is more than one item, then a rater needs to provide more than one score for each student. In other literature, "items" may be called tasks, (essay) topics, criteria (for assessing an essay), traits or constructs.

In assessing rater effects, Myford and Wolfe (2003) provided a comprehensive list of criteria including effects of leniency/severity, halo, central tendency, restriction-of-range, inaccuracy and others. Leniency/severity refers to raters' inclinations to award higher or lower scores than other raters, on average. Halo effect refers to highly correlated scores assigned across items, indicating that a rater forms a holistic view of a student's work rather than assessing each item independently of other items. Central tendency refers to the avoidance of using extreme scores so that a rater's ratings will not be outliers among

other raters. This could happen as a "play-safe" strategy on the part of a rater (Wolfe et al., 2007). Restriction-of-range is similar to central tendency, except that the limited range of scores used may be at the higher or lower end of the scale, rather than around the mean score of all raters. Accuracy has several different meanings. One concept of accuracy is about how close a rater's ratings are to some "true" scores or criterion scores. "True" scores could be derived from experts' ratings, or be the long-term average score of the rater, or the expected score based on an IRT model. In general, "accuracy" refers to whether a rater's scores for students of a particular ability is spread out (large variance), or focused around a score (small variance). A rater can be "spot-on" in terms of average leniency/severity, but be very "inaccurate" with a wide range of scores surrounding a "spot-on" average score. In contrast, an accurate rater will consistently assign similar scores to students at the same ability level.

In this paper, we will recast some of these rater effects into familiar concepts of item characteristics such as item difficulty and item discrimination. For example, there is a parallel between rater severity and item difficulty, since item difficulty reflects how many students obtain a high score, and rater severity reflects how many students are given a high score by a rater. Item discrimination refers to the extent to which an item can separate students of low and high abilities. Similarly, the range of scores a rater uses has an impact on the extent to which the rater can separate students by their abilities. That is, in this paper we will demonstrate how conventional item statistics in measurement can be applied to assess rater characteristics.

In terms of statistical models for analyzing rater effects, most analyses of rater effects involve decomposing the observed score, or a transformation of the observed score, into factors (or facets), separating rater effects from item and person (candidate/student) effects. Some analyses use raw scores as the dependent variable in a regression analysis where the observed raw score is the sum of a number of variables plus an error term. For example, Raymond and Viswesvaran (1993) modelled the observed ratings as

$$y_{nr} = \alpha_n + \rho_r + e_{nr} \tag{1}$$

where y_{nr} is the observed score for student n given by rater r, α_n is the true score for student n, ρ_r is the rater severity measure for rater r, and e_{nr} is the random error term. Variance components analyses are carried out to determine the proportion of the total variance due to different rater severity measures, different candidate abilities, and due to random errors. From these variance components, one can draw conclusions about the magnitudes of rater effects and random errors with respect to the total variance.

Other analyses use an IRT (item response theory) approach modelling the probabilities of observed scores based on student ability, item difficulty and rater severity measures. Essentially, the observed raw scores are transformed using a logit link function, and the transformed variable is decomposed into components of item difficulty, person ability, rater severity and possibly other variables. For example, a simple logit link function for a dichotomously scored item could be

$$\ln\left(\frac{p}{1-p}\right) = \theta_n - \delta_i - \rho_r \quad (2)$$

where p is the probability of success on item i, θ_n is the ability of student n, δ_i is the difficulty of item i, ρ_r is the severity of rater r. An extension of this model can be used for polytomously scored items. Additional terms can also be added to the right-hand side of Eq.(2) to include an interaction term between a rater and an item, for example. These models are commonly called the facets model (Linacre, 1989) where rater severity is a facet that has an impact on students' scores. One advantage of using a logit link function is that measures for item difficulty, student ability and rater severity are not bounded as the range of raw scores is. The second advantage of using an IRT approach is that there is usually an incomplete design that each essay is not marked by all raters, so there is a considerable amount of "missing data" in terms of a student/rater matrix. IRT approaches can handle such incomplete designs with relative ease. However, the main findings on the relative magnitudes of rater effects should be very similar between the approaches of using raw scores and using transformed scores.

In this paper, an IRT approach is used. We will discuss how common IRT models and conventional IRT parameters can be used to check on rater effects. In particular, we will use the one-parameter rating scale model (Andrich, 1978), partial credit model (Masters, 1982), and the two-parameter generalized partial credit model (Muraki, 1992). While most of these models are commonly used IRT models, we emphasise on the interpretations of standard IRT analysis results in relation to rater effects.

The case of one essay (or, one item)

For many data sets, there is only one essay being marked by raters and one overall score for an essay is assigned by a rater on each essay. For data sets where there are multiple essays written by each student, we may still be interested in analyzing each essay separately. Therefore, we will first discuss a simple way to analyse rater effect when there is only one item.

A common structure of student item response data is a two-dimensional matrix where the rows represent students and the columns represent items as shown in Figure 1.

	Item 1	Item 2	Item 3	Item 4	…
Student 1	3	2	0	2	
Student 2	1	2	1	3	
Student 3	4	3	5	4	
…					

Figure 1:
A common structure of student item response data

When there is only one item (i.e. one essay), but each student is marked by multiple raters, the structure of the student response data is shown in Figure 2.

	Rater 1	Rater 2	Rater 3	Rater 4	…
Student 1	2			4	
Student 2		3	4		
Student 3		3		4	
…					

Figure 2:
Student response data with one item and multiple raters

Figure 2 looks very similar to Figure 1 as a two-dimensional matrix except that there are some missing responses. An item response model with student and item parameters as variables in the model can be used to analyse rater data as shown in Figure 2 without involving specific facet terms (see an example rater analysis for one item in Wolfe and McVay, 2012). That is, the item parameters are now interpreted as rater parameters. This simplifies the analysis considerably since common IRT models and software programs can be used. In this respect, the commonly used item parameters of item difficulty and item discrimination can be used to interpretation rater severity and rater discrimination. As a result, in this section we will use the term "item" and "rater" interchangeably.

Item difficulty now refers to rater severity. It refers to whether a rater, on average, awards higher or lower scores than other raters. Item discrimination refers to how well a rater uses the score range to separate students. A highly discriminating rater will use low scores for low ability students, and high scores for high ability students, providing clear discrimination between low and high ability students. In contrast, a low discriminating rater may assign a particular score to students from a wide range of ability levels.

To show a practical example, we analyse a real data set of essay scores for a high-stake university entrance examination where each essay is marked by two raters. The score range is between 0 and 8. There are 886 students and 20 raters. An excerpt of the data is shown in Table 1.

The second and third columns of Table 1 are rater IDs, and the last two columns of Table 1 are the scores given by the two raters. The data is re-arranged into a two-dimensional matrix similar to that shown in Figure 2.

Table 1:
Excerpt of a data set: Scores for each student awarded by two raters

Student	Rater 1	Rater 2	Rater 1 score	Rater 2 score
1	38536	00255	0	0
2	22322	90022	4	5
3	33113	42090	7	6
4	11239	66532	2	5
5	01884	25181	1	1
6	38536	00255	3	5
7	98766	23856	4	6
8	92060	31256	3	5
…	…	…	…	…

The rating scale model and rater severity

To estimate rater severity, a rating scale model (Andrich, 1978) for polytomous item responses is fitted to the data:

$$\ln\left(\frac{p_{nik}}{p_{ni(k-1)}}\right) = \theta_n - \delta_i - \tau_k \tag{3}$$

where p_{nik} is the probability of obtaining score k for person n on item i (i.e., rated by rater i). In this model, δ_i is the item difficulty, and it represents rater severity for rater i in this example. Eq. (3) models a different rater severity parameter for each rater, but assumes the same item category threshold structure (τ_k) across raters, and the same rater discrimination. If raters do not have the same item category threshold values (τ_k), each τ_k estimated in the rating scale model represents an average across all raters. Similarly, the constant discrimination parameter (assumed to be 1 in this case) represents the average discrimination across raters. This model has the advantage of clearly defining item difficulty (rater severity), since that is the only parameter that varies across raters. The estimates of rater severity are shown in Table 2, arranged in order of severity measures.

Table 2 shows that the most severe rater is rater 14 (17322), while the most lenient rater is rater 20 (42090). To interpret the magnitudes of the range of rater severity, expected scores curves for all raters are plotted on the same graph. Each curve in Figure 3 shows the expected score given by a rater at an ability level.

It can be seen from Figure 3 that the difference between the most severe and most lenient raters is more than one score point on a 0-8 scale. For a high-stake test, this is quite a large difference.

Table 2:
Rater severity estimates from a rating scale model

Rater number	Rater ID	Rater severity (logits)
20	42090	-0.47
11	23856	-0.24
15	26484	-0.18
19	71148	-0.18
8	66532	-0.16
6	98766	-0.13
2	22322	-0.12
16	88002	-0.10
4	11239	-0.02
9	00255	0.16
13	25181	0.19
18	31256	0.25
3	33113	0.38
5	01884	0.51
17	90022	0.52
10	33908	0.54
7	92060	0.57
12	66700	0.60
1	38536	0.72
14	17322	0.88

The partial credit model and rater central tendency

The rating scale model is very restrictive in that item category thresholds (τ_k) are assumed to be the same for all raters, and raters differ only in their overall severity measures (δ_i), leading to "parallel" curves in Figure 3. A less restricted model, the partial credit model (Masters, 1982), is fitted to the data to allow the estimation of different item category thresholds for each rater. Eq.(4) shows the partial credit model.

$$\ln\left(\frac{p_{nik}}{p_{ni(k-1)}}\right) = \theta_n - \delta_{ik} \tag{4}$$

where the item category thresholds, δ_{ik} , are estimated for each rater, *i*, and score category *k*. Under the partial credit model, item difficulty is not a clearly defined notion. A rater may have high values for some δ_{ik} but at the same time low values for other δ_{ik} . Item difficulty (i.e., rater severity, in this case) is not captured by a single parameter as for the case of the rating scale model. Although, the average of the thresholds (δ_{ik}) across item categories can possibly indicate a rater's overall severity or leniency. As a result, the estimated δ_{ik} are not directly compared across raters in this paper. Instead, a plot of expected scores curve for each rater is shown in Figure 4.

The expected scores curves in Figure 4 again show a band of around one score point width, indicating that raters differ in their severity measures. Further, the curves are not all parallel: some are steeper than others. To interpret the slopes of the expected scores curves, we first note that the estimated values of δ_{ik} are dependent on the number of students in each response category, since for the Rasch model, the number of responses in each category are sufficient statistics for the parameters, δ_{ik} . That is, the shape of the

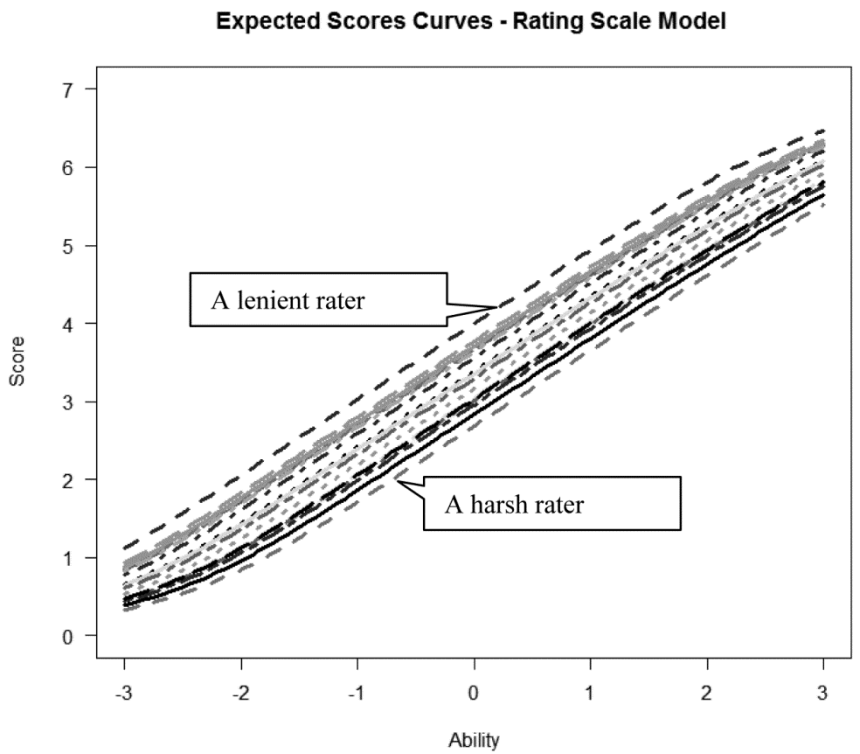

Figure 3:
Expected scores curves for raters from fitting a rating scale model

expected scores curve is determined by the frequencies of responses in each score category. As an illustration, 5000 students' responses are simulated on five items with four score categories fitting the partial credit model. Table 3 shows the number of respondents in each score category and the generating δ_{ik} .

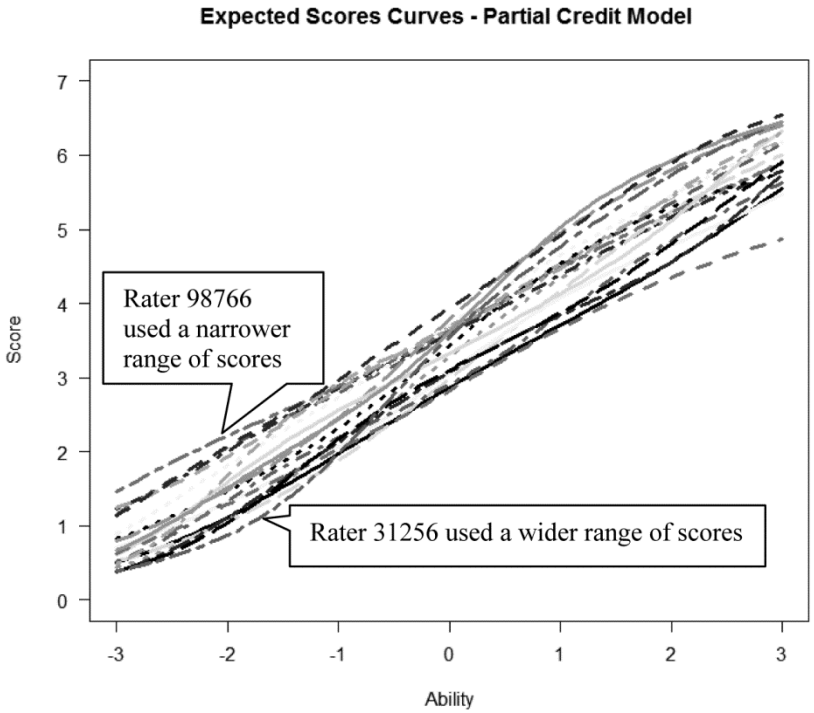

Figure 4:
Expected scores curves for raters from fitting a partial credit model

Table 3:
Simulated data: Generating PCM thresholds and frequencies of respondents in each score category

	Item 1	Item 2	Item 3	Item 4	Item 5
Generating δ_{ik}	(-2,0,2)	(-1,0,1)	(0,0,0)	(1,0,-1)	(2,0,-2)
Score 0 freq.	586	994	1547	2035	2298
Score 1 freq.	1938	1558	921	471	195
Score 2 freq.	1960	1453	941	482	202
Score 3 freq.	516	995	1591	2012	2305

Figure 5 shows the corresponding expected scores curves.

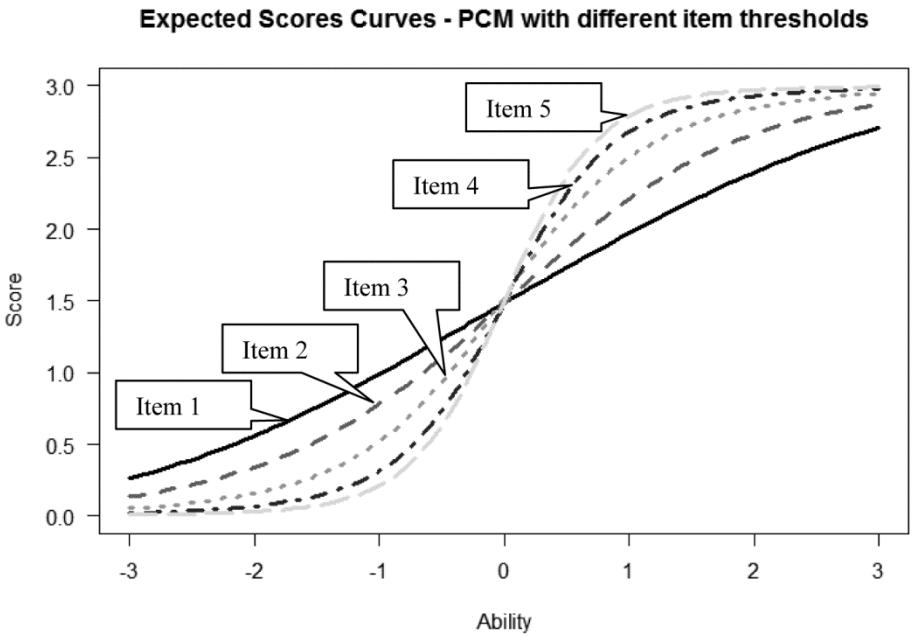

Figure 5:
Expected scores curves of five PCM items with different thresholds

From Figure 5 and Table 3, it can be seen that steeper expected scores curves are associated with more respondents in extreme score categories, while flatter curves are associated with more respondents in the middle score categories. Incidentally, the steeper expected scores curves have dis-ordered thresholds, while the flatter curves have ordered thresholds that are further apart. It is really important to note that the steepness of an expected scores curve is not an indication of the discriminating power of the item, since the partial credit model stipulates that all items (with the same maximum score) have the same discriminating power (when items fit the model, of course). Rather, the five curves in Figure 5 show that each item has different discriminating power at different points along the ability continuum. But the overall discrimination power for an item, or item information, for all five items is the same.

Going back to the real data set analysed earlier, the different slopes of the expected scores curves in Figure 4 reflect that some raters used score categories in the middle of the score range avoiding extreme scores, while others used scores at the very low and high ends of the score range. That is, the slopes of the expected scores curves can reflect the central tendency effect of raters as discussed in the introduction. Song and Wolfe

(2015) also mentioned that rater centrality can be detected in the standard deviation of raters' PCM category thresholds in a partial credit model.

We note that a simple tabulation of the frequencies of respondents in each score category marked by each rater is not so useful in this case because different raters marked different sets of essays so we do not know whether the average ability of students marked by each rater is the same. If one rater's scores are all relatively high, it could be because the rater is lenient, or it could be that the students marked by this rater are mostly of high ability.

The generalized partial credit model and rater discrimination

The partial credit model can inform us of whether raters have central tendency, but it does not provide information on rater discrimination or rater random error (inaccuracy). The thresholds of the partial credit model are determined by the number of respondents in each score category, but it does not take into account *who* are getting high scores and who are getting low scores. If two raters have identical scores distribution (i.e., with the same proportion of respondents in each score category), but the first rater assigned many high ability students low scores and assigned low ability students high scores, while the second rater appropriately assigned high scores to high ability students, the estimated score thresholds will be the same for both raters, since they have assigned the same proportion of students in each score category (the concept of raw scores being sufficient statistics for the PCM parameters). The fit statistics, however, will show large misfits for the first rater.

The term "random error" refers to the extent to which a rating departs from the expected score (for the rater). Note that in the IRT model, response categories are probabilistic. That is, it is expected that there will be variations of scores given a particular ability and rater severity. Figure 6 shows an example using simulated data. In this data set, a rater rating a student at ability level 0 has an average score of 2.5. However, the rater may give scores between 0 and 5, with a probability of 0.02 for score 0, probability of 0.12 for score 1, probability of 0.34 for score 2, etc. These probabilities are computed from fitting a partial credit model and represent the expected variation of scores under the IRT model.

However, the variation of scores for a rater may be larger or smaller than that expected by the model. A rater with large random errors is likely to have a probability distribution of scores more spread out than that for the average rater. A rater with small random errors will have a probability distribution of scores more clustered near score 2 and score 3. Further, because of ceiling and floor effects of the score range, a rater with large random errors is likely to give higher scores for low ability students, and lower scores for high ability students, thus he/she is more likely to be a less discriminating rater.

One way to check for raters' random errors is to estimate a discrimination parameter for each rater. When a rater has small random errors, the rater will have a high discrimination parameter. On the other hand, if a rater frequently assigns "incorrect" scores, his/her estimated discrimination will be low, irrespective of whether the rater is lenient or harsh.

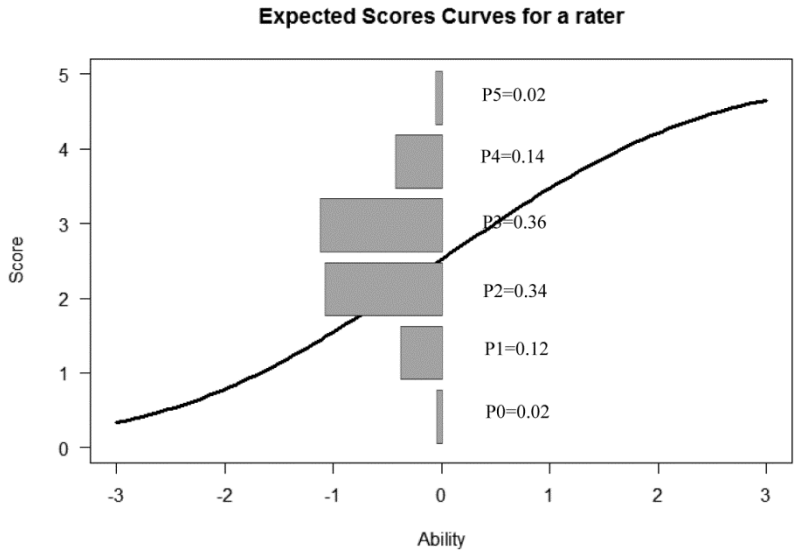

Figure 6:
Probabilities of a rater's scores for students at ability level 0

For the estimation of the discrimination parameter, the generalized partial credit model (GPCM) (Muraki, 1992) is used. Eq. (5) shows the GPCM.

$$\ln\left(\frac{p_{nik}}{p_{ni(k-1)}}\right) = a_i\left(\theta_n - \delta_{ik}\right) \tag{5}$$

where a_i is the discrimination parameter for item *i* (rater *i*, in this case).

Fitting the GPCM to the essay data analysed earlier, the 20 raters are found to have quite different discrimination parameters. Table 4 shows the estimated discrimination (a_i) parameters for the raters. Four raters (3, 11, 12, 13) have very high discrimination parameters. Unfortunately, there is no information on the background of the raters to check whether the high discriminating raters are expert/experienced raters.

A highly discriminating rater will be able to clearly separate students into low to high ability groups. A helpful way to compare high and low discriminating raters is to examine their category characteristic curves. **Figure 7** shows the category characteristic curves for four raters, ranging from highly discriminating to poorly discriminating raters.

It can be seen from **Figure 7** that the category characteristic curves for a highly discriminating rater have narrow and peaked curves, indicating that the rater assigns each score to a small ability range of students. In contrast, the category characteristic curve for a poorly discriminating rater has wide and low curves, indicating that the rater assigns each score to students of a wide range of ability levels.

Table 4:
Estimated rater discrimination parameters

Rater number	Rater ID	Rater discrimination parameter
8	66532	0.85
6	98766	0.87
10	33908	0.90
19	71148	1.38
2	22322	1.46
7	92060	1.75
15	26484	1.93
9	00255	1.94
4	11239	2.10
20	42090	2.19
16	88002	2.21
14	17322	2.52
1	38536	3.35
18	31256	3.55
5	01884	4.35
17	90022	4.47
3	33113	15.20
13	25181	19.39
12	66700	20.24
11	23856	24.07

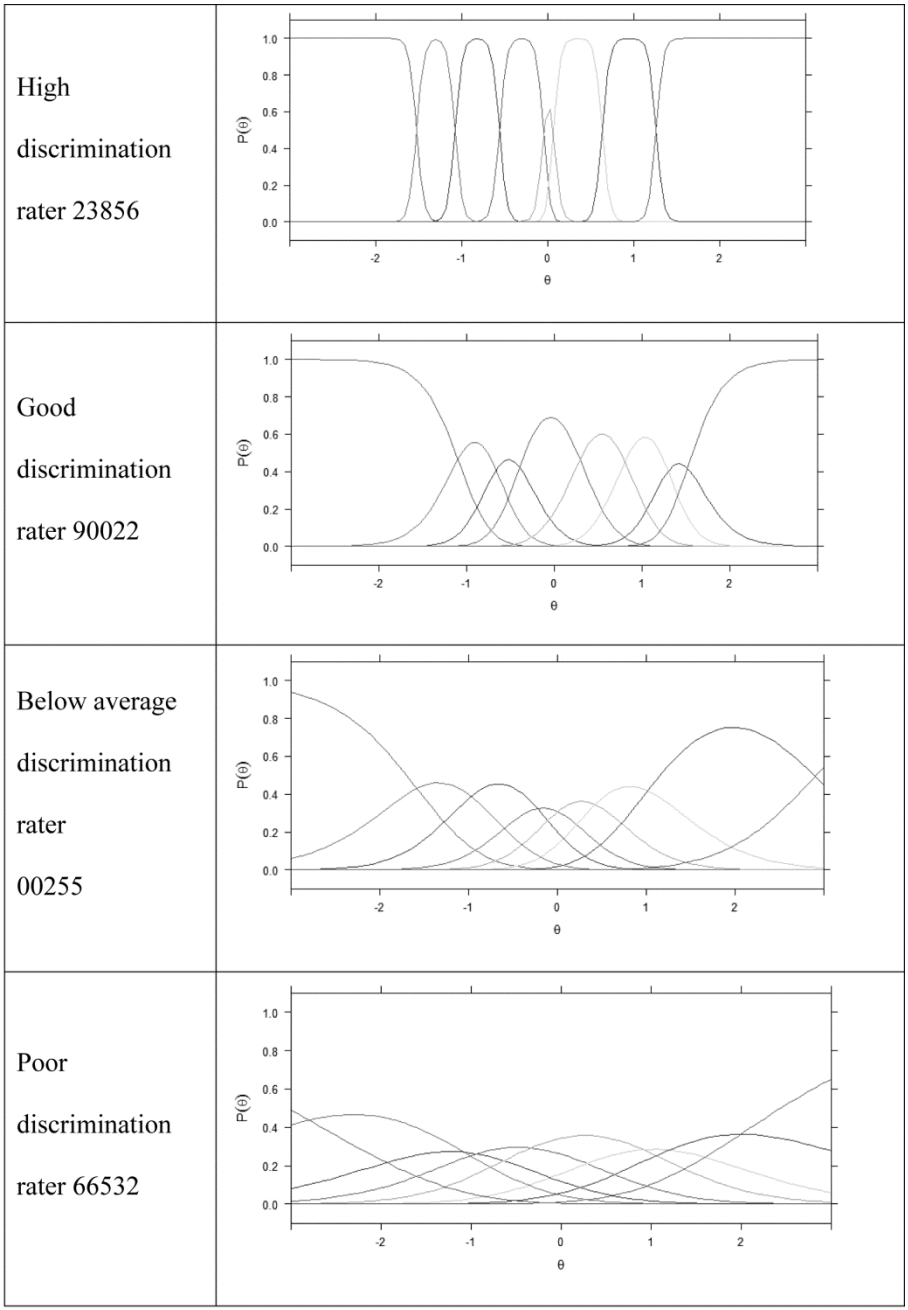

Figure 7:
Category characteristic curves for four raters

Discussion

In this paper, three item response models are used to describe three rater characteristics: rater severity, rater central tendency and rater discrimination. First, it is noted that these rater characteristics are distinct concepts and they may be quite independent of each other. A lenient rater may be a very discriminating rater, as in the case of rater 11 (23856). A plot of rater severity measures (Table 2) and rater discrimination parameters (Table 4) shows little correlation between these two, as shown in Figure 8. If the four high discrimination points are removed, the correlation between severity and discrimination is only 0.09.

A rater who uses the full range of scores may not be the most discriminating rater, as many scores may be assigned inappropriately. For example, rater 19 (71148) has used the full score range but has a low discrimination parameter. In other words, this rater has a somewhat steep expected scores curve when partial credit model is fitted. But when the GPCM is fitted, this rater is found to have quite a low discrimination parameter. In contrast, rater 14 (17322) is the most severe rater who awarded very few top two score points (a "restricted-range" rater), but he/she has a moderate discrimination parameter. Despite these cases, one may conjecture that a highly discriminating rater is likely to use the full range of scores, as it will be difficult to separate students with fewer score points.

Of the three rater characteristics: severity, central tendency, and discrimination, we place an emphasis on discrimination. The main purpose of most examinations is to assess

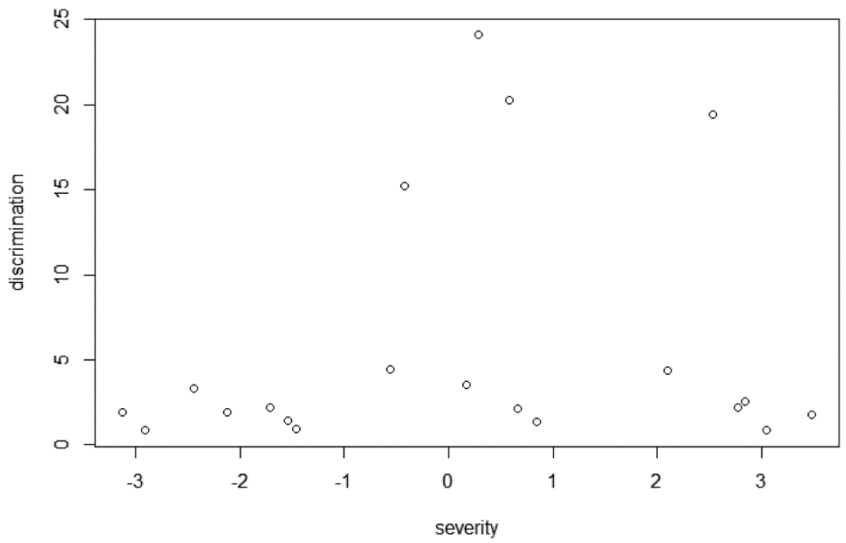

Figure 8:
Plot of raters' severity measures against discrimination measures

students' ability levels. Therefore, a good test should have high reliability and can clearly separate students into different ability levels. Adjustments for rater severity can be made when estimating student ability, provided that a rater is consistently lenient or harsh. So rater severity does not pose a significant threat to test reliability and validity. If a rater's severity is inconsistent, it will be detected through the discrimination parameter. Yet, many rater monitoring programs focus on checking severity only by comparing each rater's ratings with the scores of other raters. A very discriminating rater will likely assign quite different scores from other raters, by giving low ability students lower scores and high ability students higher scores. Consequently, if rater severity is the only criterion monitored, a very discriminating rater will likely be identified as not fitting with other raters, and be subject to corrective training. Yet a very discriminating rater is actually the best rater. For the data set analysed in this paper, the test reliability is 0.858 when a partial credit model is fitted, and 0.892 when a GPCM is fitted. If all raters can be as discriminating as for the four raters (3, 11, 12, 13), the test reliability will be even higher.

Limitations

As this paper discusses methods for analyzing one essay at a time, the methodology is not as general as many other models that explicitly model raters as facets in an IRT model. However, we note that in many facets models where there is an 'item main effect', a 'rater main effect', and an 'item by rater' interaction term, the results essentially are produced for each rater and item combination. These models, while much more elegant and general, would provide similar results as analyzing each item separately. The drawback of analyzing each essay at a time is that the halo effect would not be possible to estimate, since the halo effect refers to the dependency of scores across items. If dependency across items is ignored in an analysis, the standard errors of ability estimates would be underestimated and test reliability would be inflated. A rater model such as the Hierarchical Rater Model (HRM) (Patz et al., 2002) would be appropriate to explicitly model dependencies in the data, in particular, for the dependency caused by multiple raters rating the same piece of essay. However, the advantage of analyzing one item at a time is that simple, standard IRT programs can be used without specialized programs for analyzing facets. It is also easier to interpret item statistics in the context of rater effect. The conclusions regarding relative rater severity and rater discrimination across raters should still be valid even if dependencies in the data are not explicitly modelled. A simple procedure for providing rater information may be preferred when there is ongoing monitoring of raters as the rating process is taking place.

Clearly the validity of results of any analysis depends on the extent to which the data fit the model. The rating scale model does not fit the data as the raters do not have the same threshold structures: some raters assign more scores in the middle range while others assign more high and low scores. The data do not fit the partial credit model as the raters have different discrimination parameters. Even the generalized partial credit model may not be the best fitting model as some raters may be discriminating over one part of the score range and less discriminating over another part of the score range. Thus the expected scores curves and category characteristic curves plotted may not actually reflect

the behavior of the raters since these curves are based on the model fitted. This is one point we need to constantly remind ourselves when the results are interpreted.

This paper has not dealt with the impact of rater characteristics on the residual-based fit statistics. These fit statistics are computed using standardized residuals which reflect the difference between observed score and expected score. Some simulations conducted indicate that, depending on the model fitted, rater severity, central tendency and random errors may all have an impact on the fit statistics. It will be useful to delineate different factors that influence fit statistics under different models, and use fit statistics as another source of evidence to describe a rater's characteristics. However, much work needs to be done to thoroughly examine the relationship between fit statistics and the misfitting of the data to the model.

There are, of course, many rater effects other than severity, central tendency and discrimination. For example, a rater may find it easier to discriminate between low ability students than between high ability students. After all, when a student produces very little work, it will not be too difficult to assess. But rater discrimination may decline as the score range increases. So a model that estimates different discriminations for different score categories may be needed.

Conclusions

In conclusion, we would like to draw attention to the importance of examining different aspects of raters' assignments of scores, since focusing on just one aspect can lead to very misleading decisions about the selection of raters and rater training. No one IRT model will likely be adequate for all raters. Different IRT models may produce different profiles of raters. A rater that has steep expected scores curves under the partial credit model may be found to be a non-discriminating rater under the generalized partial credit model. A rater that has central tendency may actually be more discriminating than a rater that uses the full range of scores. Most important of all is to remember that the 'average' rater may very well be the mediocre rater while an outlying rater may actually be the best rater.

References

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

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115. doi:10.1177/0265532215582283

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

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of Rater training on reliability and accuracy of mini-CEX scores: A randomized, con-

trolled trial. *Journal of General Internal Medicine*, *24*(1), 74–79. http://doi.org/10.1007/s11606-008-0842-3

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4,* 386-422.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics,27*, 341–384.

Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. *Journals of Educational Measurement. 30*(3), 253-268.

Song, T., & Wolfe, E. W. (2015). *Distinguishing several rater effects with the Rasch model.* NCME Annual Meeting, Chicago, IL, 2015. (https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/schools/019_Rater-Effects_SongWolfe_2015NCME-1.pdf, accessed 18 September 2017).

Weigle, C. S. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

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

Wolfe, E. W., Myford, C. M., Engelhard, G., Jr. & Manolo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP* R _ *English Literature and Composition Examination using benchmark essays* (Research Report 2007-2). New York, NY: The College Board (http://files.eric.ed.gov/fulltext/ED561038.pdf, accessed 4 June 2017).