Raters' Assessment Quality in Measuring Teachers' Competency in Classroom Assessment: Application of Many Facet Rasch Model

Rosyafinaz Mohamat[1*], Bambang Sumintono[2], Harris Shah Abd Hamid[3]
[1]Faculty of Education, University of Malaya, Malaysia
[2]Faculty of Education, Universitas Islam Internasional Indonesia, Indonesia
[3]Faculty of Management, Education and Humanities, University College MAIWP International, Malaysia

*Corresponding author: rosyafinazmohamat@gmail.com

Received: 15 October 2022; **Revised**: 30 October 2022;
Accepted: 08 November 2022; **Published**: 15 November 2022

To link to this article: https://doi.org/10.37134/ajatel.vol12.2.7.2022

Abstract

This study examines the raters' assessment quality when measuring teachers' competency in Classroom Assessment (CA) using the Many Facet Rasch Model (MFRM) analysis. The instrument used consists of 56 items built based on 3 main constructs: knowledge in CA, skills in CA, and attitude towards CA. The research design of this study is a quantitative method with a multi-rater approach using a questionnaire distributed to the raters. Respondents are 68 raters consisting of The Head of Mathematics and Science Department, The Head of Mathematics Panel, and the Mathematics Teacher to assess 27 ratees. The ratees involved in this study are 27 secondary school Mathematics teachers from Selangor. The results show that among the advantages of MFRM are that it can determine the severity and consistency level of the raters, also detect bias interaction between rater and ratee. Although all raters were given the same instrument, the same aspects of evaluation, and scale category, MFRM can compare the severity level for each rater individually. Furthermore, MFRM can detect measurement biases and make it easier for researchers to communicate about the research findings. MFRM has the advantage of providing complete information and contributes the understanding of the consistency analysis of the rater's judgement with quantitative evidence support. This indicates that MFRM is an alternative model suitable to overcome the limitations in Classical Test Theory (CTT) statistical models in terms of multi-rater analysis.

Keywords: *Many Facet Rasch Model, Competency, Classroom Assessment, Rater severity, Multi-rater Analysis*

INTRODUCTION

Effective and professional teaching should be the norm in the classroom. To ensure satisfactory learning, the accomplishment of learning objectives, as well as the genuine and accurate assessment of learning, it is necessary for them to possess a thorough understanding of the subject, to be made aware of the usage of practical learning approaches and strategies, and to use many tools competently and effectively (Abdullah, 2022). However, research has confirmed that teachers' assessment abilities and capabilities are lacking (Rural, 2021).

The multi-rater approach using self-assessment and peer-assessment methods raise issues regarding the reliability of the score obtained (Donnon et al., 2013). Then inter-rater reliability is critical to increasing the reliability of the measurement. Rater effects are the factors that can influence the assessment of ratee performance (Farrokhi et al., 2011). In the multi-rater method, some ratees may be judged by severe raters, and some will be judged by lenient raters. Cronbach (1990) considers this the most serious rater's error issue. Very severe or lenient raters can contribute to a rater's error in assessment (Noor Lide, 2011). The analysis approach is usually based on Classical Test Theory (CTT) which is ideal if only one rater assesses all the ratees (Nur 'Ashiqin, 2011). In CTT, the reliability will increase if only the raters give more similar agreement in their judgement (Noor Lide, 2011). By using MFRM, the reliability and validity of the performance assessment can be improved, and conclusions on the ratee's ability are more accurate (Engelhard, 1994).

The multi-rater approach using the Many Facet Rasch Model (MFRM) can detect unexpected responses to provide information on the functions of the elements involved as if there are problems for the raters to understand and use the criteria (Eckes 2015; Kudiya et al. 2018). MFRM also has the advantage of modelled the raters based on its scale definition, without having to be in line with the assessment by other raters (Bond & Fox, 2015; Engelhard & Wind, 2018). Therefore, this article aims to show the potential of more precise and detailed rater's assessment quality using the Many Facet Rasch Model (MFRM). It has several advantages to overcome the limitations of the CTT method.

LITERATURE REVIEW

1. Multi-Rater Assessment

One of the most crucial issues in education is teacher competency in classroom assessment, which happens during the teaching and learning process (Rural, 2021). The responses from teachers demonstrate the program's influence on the growth of their assessment competencies, particularly about formative and summative assessment and creating various types of assessments in line with achievement criteria (Tomasevic et al., 2021). Quality assessment teaching, also known as assessment literacy, depends on teachers' readiness to comprehend and apply data in the classroom. (Hodges et al., 2019; Seifert & Feliks, 2019).

Competency elements are aspects of knowledge, skills and attitudes that can predict individual achievement, behaviour, teacher's quality and professionalism and student innocence. The researchers built the theoretical framework for this study based on the Spencer and Spencer (1993) theory (Figure 1). Several predictor factors influence the teacher's competency level, namely the teaching professionalism value practise, training, emotions, social, cognitive intelligence, background, and self-readiness. The measurement of competencies made is related to the achievement of teachers, which affects the effectiveness of the classroom assessment.

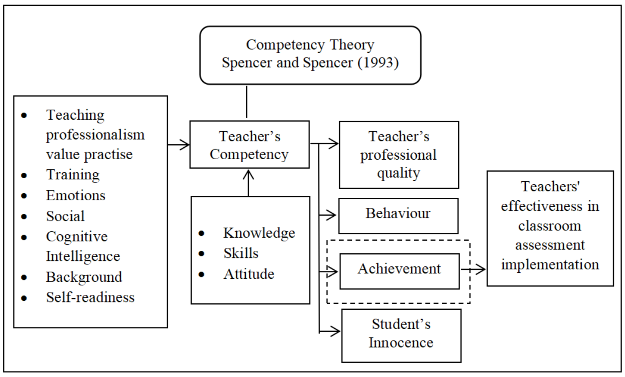


Figure 1 Theoretical Framework

Rater's assessment is usually subjective and can affect the reliability and validity of the ratee performance (Schaefer, 2008). Using a single rater can result in a biased assessment (Matsuno, 2009). To overcome this limitation, the use of self-assessment and peer-assessment has increased in the education field (Hargreaves et al., 2002). The rater effects include various unwanted phenomena, including inconsistent raters, rater severity and bias judgement, which can contribute to undesirable diversity in the measurement process (Han, 2021). In addition, the study by Sahin et al. (2016) also found that the respondents responded positively by stating that peer assessments were not complex, helping them understand their friends and enhancing their learning and self-confidence.

The multi-rater method produces a more stable and accurate assessment and has higher reliability than the self-assessment method (Calhoun et al., 2011; Goffin & Jackson, 1992; Lohman, 2004). For instance, the reliability of teacher assessment is higher when it involves more raters (Kane & Staiger, 2012). The multi-rater method has become increasingly popular involving peer-assessment, self-assessment and superiors or subordinates' assessment to determine an individual's job performance (Scullen et al., 2000). The assessment made by a colleague can enhance the reliability and validity of the evaluation made in line with the assessment of the work assignment aspect (Schmidt et al., 2016).

The previous studies show the MFRM as a proper psychometric framework compared to the Classical Test Theory (CTT) method to consider the rater effects, as the MFRM is more general and can provide a detailed analysis of raters' judgement (Eckes 2019). The MFRM is based on the judgement of unrelated raters, unidimensionality properties and same item discrimination between high-ability and low-ability ratee (Styck et al., 2020). Assessing teacher quality using performance assessment is recommended to involve more than one rater as the involvement of several raters is often seen as the 'key' to successful teacher assessment practices (OECD, 2013). The research that uses a multi-rater approach assumes that it can obtain a more accurate and fair assessment.

2. Common Method Used in the Multi-Rater Analysis

Various methods have been widely used to determine the consistency of raters based on the CTT approach. For example, the Cohen Kappa method measures the consistency between two raters by excluding the agreement between the two raters (Hsu & Field, 2003). Next, the Fleiss Kappa method provides statistical comparison interpretations that are easier to understand than the Cohen Kappa method, which is more difficult to interpret the determination of the rater's agreement (Allen, 2017). The following tool is Generalizability Theory (G theory), developed by Lee Cronbach to measure the reliability between raters and has the advantages of isolating and assuming the various sources (Brennan, 2010; Webb et al., 2018). The G theory is an extended statistical theory of CTT that allows a more precise calculation of reliability related to the behavioural measurements and can assume the various error's sources to calculate the reliability more precisely (Nor Mashitah, 2017). Content Validity Index (CVI) is another method that can be used to determine the validity of the overall content of the instrument in multi-rater situations, calculated based on the average Content Validity Ratio (CVR) (Lindell & Brandt, 1999). CVI provides direct information on the rater's agreement by converting ordinal scale data into two categories (example: relevant or irrelevant) (Polit & Beck, 2006).

3. Weaknesses of Existing Methods

There are various disadvantages in multi-rater analysis methods by using the CTT approach. The Cohen Kappa method can be used if the total number of raters is two, while Fleiss Kappa can be used for more than two raters, but only with nominal data categories (Cohen, 1960; Fleiss & Cohen, 1973). However, the Fleiss Kappa method is questionable because it depends on the assumption of homogeneity and is difficult to use for polytomous data (Allen, 2017; Bartok & Burzler, 2020; Warrens, 2010). The Fleiss Kappa method is also unable to detect if there is a possibility of guessing performed by the raters in the scoring process and is unable to detect the severity level of the raters (Allen, 2017).

In addition, the internal consistency measurement based on CTT has a limitation because it cannot systematically distinguish the raters, for example, when the severity level of the raters is consistent with all ratees (Newton, 2009). Although G Theory has several advantages over the commonly used CTT method, it is quite complex and complicated, making it difficult for the reader to accept and understand the interpretation (Brennan, 2010; Webb et al., 2018). The G Theory also has some limitations, such as not determining the severity level of the raters and causing the rater's error cannot be included in the explanation of the scale testing (Zhu et al., 1998).

Furthermore, the CVI method also has several limitations, such as involving only two categories of ordinal scale, the rater's agreement index is likely to decrease if the number of raters increases, using the average value approach to determine the rater's agreement, and only focusing on item suitability but not involving scale analysis to ensure the construct measurements were made accurately (Polit & Beck, 2006). The CVR method is only limited to assessments for dichotomous data (Lindell & Brandt, 1999).

4. MFRM in Research

One of the advantages of using the Rasch measurement model is that this model can estimate the individual's abilities without relying on the item and the estimated item parameters are also free without relying on individual groups (Sumintono, 2016). MFRM is an advanced Rasch measurement model and involves more than two interacted aspects to produce observation (Linacre, 1994). MFRM can combine more facets to determine the relationship between the facets, for example, an analysis involving three facets, i.e., items, raters and ratees (Eckes, 2015). In the comparison of the rater's judgement, MFRM can explain clearly the severity level of the raters, the consistency of the raters, correcting the rater's score based on the ideal model, rating scale analysis and investigating bias interactions (Bond & Fox, 2015; Eckes, 2015; Engelhard & Wind, 2018). A study by Cai (2015) showed that biased judgment could affect the assessment process in the tests.

The analysis by using MFRM has gained much attention from researchers and has been widely used in language testing, education and psychological measurement (Barkaoui, 2013; Linacre, 1994). MFRM is also widely used in other areas such as study in nutrition by Sunjaya et al. (2020), research to determine the quality of rater's judgement in The Canadian English Language Benchmark Assessment for Nurses (CELBAN) by Wang et al. (2021) and research to analyse the content validity for Computerized Testlet Instrument to Measure Chemical Literacy Capabilities by Fahmina et al. (2019). MFRM also has advantages compared to CTT because MFRM can identify inaccurate responses by the raters, inappropriate judgement patterns, and detect missing data (Fahmina et al., 2019; Goodwin & Leech, 2003). MFRM can detect biases in measurements and make it easier for researchers to communicate about the research findings (Boone, 2020). MFRM contributes to understanding consistency analysis of rater's judgement with quantitative evidence support (Nor Mashitah et al., 2015; Zhu et al., 1998).

METHODOLOGY

1. Instrument

The instrument used in this research measures teachers' competency in Classroom Assessment (CA). This instrument consists of 56 items that are built based on three main constructs, namely knowledge in CA (22 items), skills in CA (24 items) and attitude towards CA (10 items). The instrument determination constructs are based on the analysis of 8 competency models and 13 existing competency instruments, adjusted to the Classroom Assessment Implementation Guidelines (Second Edition) from Bahagian Pembangunan Kurikulum (2019). The raters will respond to all items to measure the ratee's competency in CA. Each item was assessed based on a 5-point Likert scale as response options for all the items; the higher the score, the better the performance of the ratee.

2. The Respondents

The sample of this study was Mathematics teachers, where the total number of teachers as ratee involved in this study is 27, and there were 68 raters recruited to assess these teachers. Each teacher (ratee) is rated only by four raters, and each raters assessed more than two teachers in many cases. Therefore, the total number of responses collected in this study is 108 (27 ratees × 4 raters). The background of raters who assessed a teacher consists of different backgrounds; detail of their demographic is shown in Table 1.

Table 1 Background information of the raters (N = 68)

Demographic	Factors	Frequency	Percent (%)
Gender	Male	5	7.35
	Female	63	92.65
Age	20-29 years	1	1.47
	30-39 years	36	52.94
	40-49 years	25	36.76
	50-60 years	6	8.82
Position	The Head of Mathematics & Science Department	7	10.29
	The Head of Mathematics Panel	7	10.29
	Mathematics Teacher	54	79.41
Experience	1-9 years	21	30.88
	10-19 years	40	58.82
	20-29 years	7	10.29

The population of ratee for this study are Mathematics teachers who serve in the government secondary schools in Selangor. Selangor has a large population and can represent the characteristics of Malaysia's population. Selangor has the largest number of teachers compared to other states. Apart from that, Selangor is also the state with the highest number of secondary schools after Johor. In this study, several sampling techniques were used to identify the respondents. The cluster sampling technique was used to categorise Selangor into ten districts. Then, simple random sampling was used to select four districts, two schools for each district, four teachers for each school, and four raters for each ratee.

3. Measurement Model

The collected data was analysed using MFRM to determine rater severity, consistency and bias interaction that occurs in the assessment by the raters. The fit statistics are essential to help the researchers to know the extent of accuracy of the data fit to the Rasch model (Siti Rahayah, 2008). The value of Infit MnSq and Outfit MnSq in fit statistic shows the rater's consistency in performing the assessment. The value of MnSq = 1 indicates that the data is ideal according to Rasch model specifications. The acceptable value of MnSq in fit statistic is between 0.5 to 1.5 (Bond & Fox, 2015). The reliability index for the data is accepted if the value is above 0.65 (Bond & Fox, 2015). The analysis to determine the separation index was carried out to obtain the assumptions or estimations of separation or differences of respondents based on the level of ability on the measured variables (Wright & Masters, 1982). If the separation index obtained is more than 2, it indicates a good and accepted value (Linacre, 2006). Rasch analysis requires at least a minimum of 40% raw variance explained by measures as an indicator of good unidimensionality instrument (Bond & Fox, 2015).

RESULTS

The analysis results showed that the number of responses involved was 6048 (27 ratee × 4 raters × 56 items), indicating no missing data. The data were recorded in Microsoft Excel software and then analysed using FACETS version 3.71.3, which involves three facets; raters, ratee and items.

1. Reliability and Construct Validity

To determine the reliability of the rater's assessment, the researchers looked at the value of the reliability and validity index from the MFRM analysis findings (Table 2).

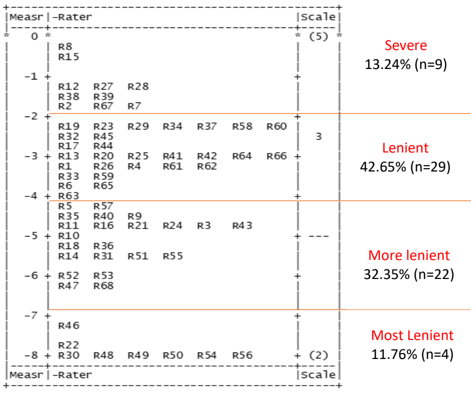
Table 2 MFRM analysis findings

	Rater	Ratee
N	68	27
Mean of logit	-4.14	0.00
standard deviation (SD)	2.36	0.87
standard error (SE)	0.45	0.11
Separation Index	3.67	3.71
Strata	5.22	5.28
Reliability Index	0.93	0.93
Significance (probability) (p)	0.00	0.00
Observed Exact Agreements (%)	59.0	
Expected Agreements (%)	57.7	
Variance explained by Rasch measures (%)	42.52	

The value of the rater's reliability index is high, which is 0.93, the separation index of 3.67 is also good as it is above 3. The significance (probability) value of p = 0.00 indicated a significant difference to the severity level of the rater, and there is a high internal consistency in the assessment by the raters. This indicated that the panel had different severity levels when doing the assessment. The two percentages of rater agreement values indicate inter-rater reliability, its shows that almost the same value indicates the data meets the expectations by the Rasch Model. In Rasch's analysis, the percentage of variance explained by Rasch measures needs to reach at least a minimum of 40% to demonstrate good unidimensionality (Engelhard & Wind, 2018). The findings showed that the analysis has good reliability and construct validity.

2. Severity Level of Rater

The logit value from Facets software indicates the rater's assessment to determine the respondent's ability level, the item's difficulty level and the rater's severity level. Wright's map helps the researchers compare individual rater severity and leniency (Boone, 2020). The mean measure (logit) of raters was -4.14 indicating all rater's tendency to give a higher score easily (lenient). However, the standard deviation value suggests a wide dispersion of measures across the raters' logit scale (SD = 2.36) which indicates the raters have a different severity level. Figure 2 informs that the position of the rater R8 at the top is the most severe rater while the position of the rater R30 below in the chart as the most lenient rater.



```
+--------------------------------------------------+
|Measr|-Rater                               |Scale|
|-----+-------------------------------------+-----|
*   0 *                                     * (5) *
|     | R8                                  |     |
|     | R15                                 |     |
|     |                                     |     |
|  -1 +                                     +     |
|     | R12  R27  R28                       |     |
|     | R38  R39                            |     |
|     | R2   R67  R7                        |     |
|  -2 +                                     +     |
|     | R19  R23  R29  R34  R37  R58  R60   |     |
|     | R32  R45                            |  3  |
|     | R17  R44                            |     |
|  -3 + R13  R20  R25  R41  R42  R64  R66   +     |
|     | R1   R26  R4   R61  R62             |     |
|     | R33  R59                            |     |
|     | R6   R65                            |     |
|  -4 + R63                                 +     |
|     | R5   R57                            |     |
|     | R35  R40  R9                        |     |
|     | R11  R16  R21  R24  R3   R43        |     |
|  -5 + R10                                 + --- |
|     | R18  R36                            |     |
|     | R14  R31  R51  R55                  |     |
|     |                                     |     |
|  -6 + R52  R53                            +     |
|     | R47  R68                            |     |
|     |                                     |     |
|     |                                     |     |
|  -7 +                                     +     |
|     | R46                                 |     |
|     |                                     |     |
|     | R22                                 |     |
|  -8 + R30  R48  R49  R50  R54  R56        + (2) |
|-----+-------------------------------------+-----|
|Measr|-Rater                               |Scale|
+--------------------------------------------------+
```

Figure 2 Wright map of rater's severity level (N=68)

Further, six raters in the most lenient groups (R30, R48, R49, R50, R54, and R56) which is 8.82% from the total, were also outliers because they were too lenient. Their demographic profile was female raters from different age groups and job positions (for instance, R50, R54, R56 and R30 in 30-39 years group and mathematics teacher). The diverse demographic characteristics among this outlier group indicate no specific identity detected for most lenient raters. The measurements will become weak if outliers are not removed (Linacre, 1994).

Figure 2 shows that most raters tended to be lenient (58 or 85%), with a logit value below -2 logit. There is a possibility that this is because most ratees being assessed are very good and have a high ability. There were only five male raters, namely R8, R33, R37, R39 and R40. The positions of male raters were scattered and not clustered. There were six raters aged 50-60 years (R43, R46, R52, R55, R58 and R59) in the three categories. These findings showed that the rater's gender and age do not affect the rater severity level.

There were seven raters with 20-29 years of experience, namely R21, R65, R52, R55, R59, R24 and R58. These seven raters were more lenient and lenient categories. This shows that raters with more experience tend to give a lenient judgement in this study. Seven raters held positions as The Head of Mathematics & Science Department, namely R2, R12, R24, R36, R42, R48 and R58. These seven raters were in all categories. This also indicates that the rater's position does not affect the rater severity level.

3. Fit statistics of Raters

The data screening process found that seven misfit raters had outfit values of its MnSq and Zstd that did not meet the acceptable range.

The findings demonstrated that seven raters (10.29%), as shown in Table 3, were misfits. Raters R7, R29, R63 and R64 have the same demographic characteristics, they are female raters, job positions as Mathematics teachers, and the age range is 30 to 39 years. Raters R34 and R62 are female raters, job positions as Mathematics teachers, and the age range is 40 to 49 years. While rater R37 is a male rater, job position as The Head of Mathematics Panel and the age range is 40-49 years. The diverse demographic characteristics among misfit raters indicate no specific identity detected for misfit raters, and any rater can be a misfit rater.

Table 3 Fit Statistics Analysis Findings

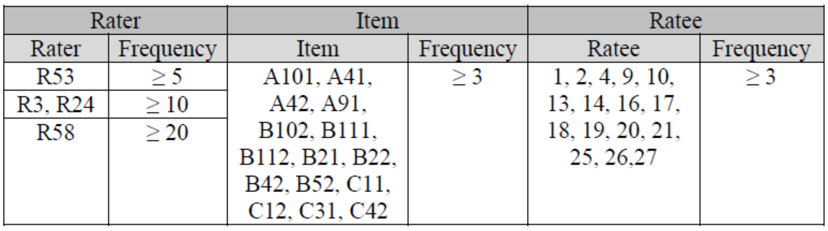
Rater	**Outfit**		**Correlation**
	MnSq	**Zstd**	**PtMea**
R7	***0.06***	***-6.42***	0.00
R29	***0.46***	***-2.81***	0.42
R34	***0.37***	***-3.30***	0.41
R37	***0.28***	***-5.46***	0.30
R62	***0.06***	***-6.42***	0.00
R63	***2.23***	***4.4***	0.29
R64	***0.06***	***-6.42***	0.00

Overall, the data screening process showed that only 13 raters (six outliers and seven misfit raters) responded differently to Rasch's ideal model, which showed a sensitive analysis from this measurement model. Although the findings indicated that most raters are fit, the researchers are also interested in studying the sensitivity of MFRM further. The following analysis stage is to identify unexpected responses and bias interaction between the rater and ratee.

The unexpected response findings indicated that MFRM could detect the consistency for each rater on a particular item (Refer Appendix C). 77 responses showed the rater gave a lower score than the expected score (under-value) and 23 responses that showed the rater gave a higher score than the expected score (over-value). The number of unexpected responses detected was too small at only 1.65% (100 out of 6048 responses), indicated that all raters had made a cautious and detailed assessment. Table 4 shows some of the unexpected responses with high frequency for the three facets (rater, item and ratee), which can provide information about the consistency of the rater and the quality of the items.

Rater R58 is less consistent because it has the highest frequency of unexpected responses. Rater 58, who has made unexpected responses, was a fit rater. This shows that the findings of unexpected responses are not direct evidence that can determine the misfit rater. Items A101, A41, A42, A91, B102, B111, B112, B21, B22, B42, B52, C11, C12, C31 and C42 (27% from total item) caused the rater to be confused when doing the judgement because they have a high frequency of unexpected responses compared to other items.

Table 4 Summary of unexpected response analysis findings



Rater		Item		Ratee	
Rater	Frequency	Item	Frequency	Ratee	Frequency
R53	≥ 5	A101, A41, A42, A91, B102, B111, B112, B21, B22, B42, B52, C11, C12, C31, C42	≥ 3	1, 2, 4, 9, 10, 13, 14, 16, 17, 18, 19, 20, 21, 25, 26,27	≥ 3
R3, R24	≥ 10				
R58	≥ 20				

Meanwhile, bias interaction occurs when a discrepancy between the observed score value and the expected score value detected based on the Rasch model's ideal model. Raters who are not consistent in their assessment tend to give a bigger observed score than the expected score or give a smaller observed score than the expected score as indicated with a Rasch-Welch t-value bigger than +2 or less than -2 (Table 5).

Table 5 Bias interaction rater-ratee

Rater	Ratee	Observed Score	Expected Score	Average O-E	Bias Measure	S. E	t value	Outfit MnSq
R25	10	211	217.52	-0.12	-0.58	0.29	-2.00	1.1
R25	11	211	227.39	-0.29	-1.57	0.29	-5.45	0.9
R25	13	265	243.28	0.39	1.71	0.31	5.53	1.3
R25	14	231	223.74	0.13	0.73	0.31	2.35	1.1
R24	11	265	248.22	0.30	1.34	0.31	4.33	0.9
R24	13	244	265.69	-0.39	-1.72	0.28	-6.23	0.9
R24	14	236	243.29	-0.13	-0.59	0.29	-2.02	1.4

The result shows the total number of bias interactions between rater and ratee is very low, only 6.48% (7 out of 108 responses). This suggests that the raters have made consistent assessments and made less mistakes. Two raters tend to have more bias in their assessment, the rater R25 (4 times) and the rater R24 (3 times). The researchers found that all raters who have made biased assessments against ratee were fit raters. This also shows that a biased assessment does not cause the misfit rater, and even a fit rater may be biased in the assessment.

The rater R24 shows leniency in assessment towards ratee 11 based on the large difference between the observed and the expected score of 16.78 points (265 – 248.22 = 16.78). Meanwhile, the rater R24 shows severity in assessment towards ratee 13 and ratee 14 based on the large difference between the observed score and the expected score for ratee 13 and ratee 14. The lenient raters gave the ratee a higher observed than the expected score with a t value above +2. The lenient raters gave the ratee a lower observed than the expected score with a t value below -2. The findings show that rater R25 and rater R24 contributed to significant bias interactions, including over-value or under-value. The Outfit MnSq values for all detected bias interactions ranged between 0.9 to 1.4, within the acceptable value.

Figure 4 shows eight misfit raters who showed bias and inconsistency in their assessments. The plot at the top of the graph shows that the rater has made a severe assessment, given a lower score. At the same time, the plot at the bottom of the graph shows that the rater has made a lenient assessment and scored higher. For example, rater R62 was severe when assessed ratee 25, but lenient when assessed ratee 24. Figure 4 also clearly shows some of the bias judgements made by rater R24 (provide a higher score to rate 11, 13 and 14) and rater R25 as mentioned in the interaction bias analysis findings, where these two raters show inconsistency.

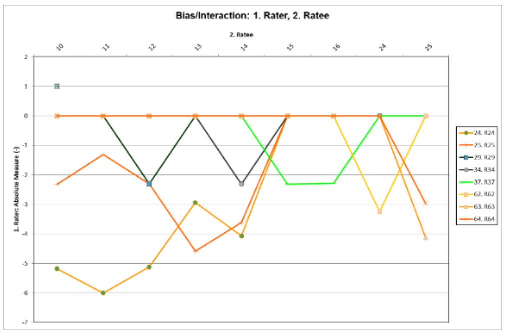


Figure 4 Bias among misfit raters

DISCUSSIONS

The data analysis in the study show that it is fit with the Rasch model (Table 2), principal component analysis of residuals is more than 40% indicating good unidimensionality of the instrument used (Andrich & Marais, 2019; Liu & Lim, 2020). This suggest that three constructs with 56 items of the instrument works very well to measure latent variable of ratees' classroom assessment with multi-rater approach (Bond & Fox, 2015; Mohd Zabidi et al., 2022). Further all reliability indices (reliability, strata and separation) showing excellent result, a kind of multi rater approach situation where volume data increase compared to self-administered data for instance (Eckes, 2015; Englehard & Wind, 2018). All in all, at the instrument level the findings showed that the MFRM could analyse the reliability and validity of the instrument thoroughly in multi-rater situations and detail compared to another measurement model (Boone et al., 2014; Eckes, 2015; Englehard & Wind, 2018).

One distinctive analysis using Rasch model is, it can provide individual-centered statistics, in this study it showed that MFRM could detect detailed information about rater severity and leniency (Engelhard & Wind, 2018). In this study, using mean and standard deviation of raters' logit, raters' severity divided into four groups and its number too (Figure 2). The result showing that raters tend to be lenient which can mean mathematics teacher being assessed has good competency (Mohd Yusri et al., 2019; Nurul Farahin & Siti Mistima, 2021), though several raters also consider as severe with strict evaluation. Identification of raters' severity and leniency level showing powerful analysis of the MFRM, something that missing from other approach (Eckes, 2015; Boone et al., 2014; Mohd Zabidi et al., 2022). There is a possibility that the rater's severity level is influenced by various factors, such as the difference of raters in terms of opinion, experience, and background knowledge about the domain being judged (Styck et al., 2020). Gender, age and amount of training received can also be the other factors that influence the rater's judgement (Eckes, 2015). Raters varied significantly in age, gender, education, which may have contributed to no significant findings between personality traits and rating severity (Zhu et al., 2021). But this study found that gender, age and position do not affect the rater severity level when judging mathematic teachers.

Further, the finding of the study also analyzed fit statistic raters which informing their quality work. The diverse demographic characteristics among misfit raters indicated that no specific identity was detected, and any rater can be a misfit rater; showing sensitivity of individual centered statistics analysis (Eckes, 2015; Mohd Zabidi et al., 2022). In addition, the diverse demographic characteristics found among the outliers group indicated no specific identity detected for most lenient raters; indicating MFRM has advantages in providing information to the individual level (Engelhard & Wind, 2018).
Other useful analysis of MFRM is it can detect inconsistency of raters in terms of unexpected response and biased assessments. The findings detected 100 unexpected responses, which was 1.65% from total showing most raters conducted their assessment professionally. Regarding bias, the findings also showed that there are 2 biased raters, namely rater R25 with 4 bias interactions and rater R24 with 3 bias interactions. However, this study found that unexpected responses and biased interactions could not support the misfit information. A study conducted by Sunjaya et al. (2020) also showed the ability of MFRM to detect 15 unexpected responses that can explain the consistency of the rater's judgement. The findings on bias interaction and unexpected responses showed the advantages of MFRM to provide

evidence regarding multi-rater quality assessment and ensure the measurement is produced more accurate and precise (Andrich & Marais, 2019; Bond & Fox, 2015).

MFRM can also help researchers identify the rater's demographic information from each severity group. These advantages are essential to obtain a fair and precise assessment based on the rater's judgement (Eckes, 2019; McNamara & Knoch, 2012). The research findings by Springer and Bradley (2018) showed specific observed trends that cannot be detected by using the CTT approach, like finding of this present study. The multi-rater analysis methods using CTT, such as Cohen Kappa, Fleiss Kappa and G Theory, have some limitations, such as cannot determining the severity level of the raters, bias judgement and unexpected responses. The rater's consistency analysis showed that the raters had a different severity level when judging and empirical evidence on the analysis of bias interactions (Schaefer, 2008). The analysis conducted can determine the severity level of the raters and improve the validity of the process (Mohd Zabidi et al., 2022). The researchers can reflect on the diversity of raters affecting judgement results (Eckes, 2019; Fan et al., 2019). Other than that, research by Schaefer (2008) found that there were raters who rated higher ability ratee very severely, and there were also raters who rated lower ability ratee very leniently. As show in the present study, unexpected response and bias can be detected with MFRM, which implicated better analysis can be resulted (Engelhard & Wind, 2018).

As indicated in other studies (Lumley & Mcnamara, 1995; Shin, 2010; Wigglesworth, 1993), raters training is needed in order to improve the rater's consistency; whereas unexpected response and bias interaction as evidences show in this present study. Analysis using FACETS in this research can be used as feedback about the rater and the rater's behaviour to a particular task (Eckes, 2015). The use of FACETS can explain the findings of the rater bias into a rater training program so that the raters can be aware of their behaviour and severity level to improve their consistency in the judgement.

CONCLUSION

The findings of multi-rater methods analysis by using MFRM shows exciting results and comprehensive information on the consistency of the raters. MFRM can be used to identify and avoid biased judgment, identify the poor-quality raters and detect the bias interactions in the assessment. This study also shows that measuring teacher competency is not easy, but MFRM is an excellent tool to identify it. Unlike the CTT's approach that emphasises on group-centered statistics, MFRM produces more detailed information on the pattern of the rater's tendencies, the rater's severity level and improving the validity process (Mohd Zabidi et al., 2021). Overall, the study showed MFRM as an effective psychometric framework compared to CTT's method, investigating the rater effects because MFRM is more general and provides a detailed analysis of the rater's assessment (Eckes, 2019).

ACKNOWLEDGEMENT

The author would like to express his deepest gratitude to all parties for the support .

FUNDING

The authors declare that no financial support was received for the research of this article.

DATA AVAILABILITY

Data will be made available on request.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

Allen, M. (2017). *The SAGE encyclopedia of communication research methods* (Volume 1). Retrieved from United States of America

Abdullah Al-Awaid, S. A. (2022). Online education and assessment: Profiling EFL teachers' competency in Saudi Arabia. *World Journal of English Language*, *12*(2), 82. https://doi.org/10.5430/wjel.v12n2p82

Bahagian Pembangunan Kurikulum. (2019). *Panduan pelaksanaan pentaksiran bilik darjah edisi Ke-2*. Putrajaya: Kementerian Pendidikan Malaysia.

Barkaoui, K. (2013). Multifaceted Rasch analysis for test evaluation. *The Companion to Language Assessment*, 1–46. https://doi.org/10.1002/9781118411360.wbcla070

Bartok, L., & Burzler, M. A. (2020). How to assess rater rankings? A theoretical and a simulation approach using the sum of the Pairwise Absolute Row Differences (PARDs). *Journal of Statistical Theory and Practice*, *14*(37). https://doi.org/10.1007/s42519-020-00103-w

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (Third Edit). New York: Routledge Taylor & Francis Group.

Boone, W. J. (2020). Rasch basics for the novice. In *Rasch measurement: Applications in quantitative educational research* (pp. 9–30). Singapore: Springer Nature Singapore Pte Ltd.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly*, *12*(3), 262–282. https://doi.org/10.1080/15434303.2015.1053134

Calhoun, A. W., Boone, M., Miller, K. H., Taulbee, R. L., Montgomery, V. L., & Boland, K. (2011). A multirater instrument for the assessment of simulated pediatric crises. *Journal of Graduate Medical Education*, *3*(1), 88–94.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Cronbach, L. J. (1990). *Essentials of Pychological Testing* (5th Editio). New York: Harper & Row.

Donnon, T., McIlwrick, J., & Woloschuk, W. (2013). Investigating the reliability and validity of self and peer assessment to measure medical students' professional competencies. *Creative Education*, *4*(6), 23–28. https://doi.org/10.4236/ce.2013.46a005

Eckes, T. (2015). *Introduction to Many-Facet Rasch measurement: Analyzing and evaluating rater-mediated assessment*. Frankfurt: Peter Lang Edition.

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative Data Analysis for Language Assessment Volume I: Fundamental Techniques* (pp. 153–176). https://doi.org/10.4324/9781315187815-2

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G., & Wind, S. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York: Routledge Taylor & Francis Group.

Fahmina, S. S., Masykuri, M., Ramadhani, D. G., & Yamtinah, S. (2019). Content validity uses Rasch model on computerized testlet instrument to measure chemical literacy capabilities. *AIP Conference Proceedings*, *2194*(020023). https://doi.org/10.1063/1.5139755

Fan, J., Knoch, U., & Bond, T. G. (2019). Application of Rasch measurement theory in language assessment: Using measurement to enhance language assessment research and practice. *Papers in Language Testing and …*, *8*(2).

Farrokhi, F., Esfandiari, R., & Dalili, M. V. (2011). Applying the many-facet Rasch model to detect centrality in self-assessment , peer-assessment and teacher assessment. *World Applied Sciences Journal*, *15*, 70–77.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*, 613–619. https://doi.org/10.1177/001316447303300309

Goffin, R. D., & Jackson, D. N. (1992). Analysis of multitrait-multirater performance appraisal data : Composite direct product method versus confirmatory factor analysis. *Multivariate Behavioral Research*, *27*(3), 363–385.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing. *Measurement and Evaluation in Counseling and Development*, *36*(3), 181–191.

https://doi.org/10.1080/07481756.2003.11909741

Han, C. (2021). Detecting and measuring rater effects in interpreting assessment: A methodological comparison of classical test theory, generalizability theory, and many-facet Rasch measurement. *New Frontiers in Translation Studies*, (April), 85–113. https://doi.org/10.1007/978-981-15-8554-8_5

Hargreaves, A., Earl, L., & Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal*, *39*(1), 69–95.

Hodges, T. S., Scott, C. E., Washburn, E. K., Matthews, S. D., & Gould, C. (2019). Developing pre-service teachers' critical thinking and assessment skills with reflective writing. In *Handbook of Research on Critical Thinking Strategies in Pre-Service Learning Environments* (pp. 146-173). IGI Global. https://doi.org/10.4018/978-1-5225-7823-9.ch008

Hsu, L. M., & Field, R. (2003). Interrater agreement measures: Comments on Kappa n , Cohen's Kappa, Scott's π, and Aickin's α. *Understanding Statistics*, *2*(3), 205–219. https://doi.org/10.1207/s15328031us0203_03

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Washington: Bill and Melinda Gates Foundation.

Kudiya, K., Sumintono, B., Sabana, S., & Sachari, A. (2018). Batik artisans' judgement of batik wax quality and its criteria: An application of the many-facets Rasch model. In Q. Zhang (Ed.), *Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings* (pp. 27–38). https://doi.org/10.1007/978-981-10-8138-5

Linacre, J. M. (1994). *Many-facet Rasch Measurement*. Chicago: MESA PRESS.

Linacre, J. M. (2006). *A user's guide to Winsteps/ Ministep Rasch-model computer programs*. Chicago: www.winsteps.com.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the cvi, t, rwg(j), and r*wg(j) indexes. *Journal of Applied Psychology*, *84*(4), 640–647. https://doi.org/10.1037/0021-9010.84.4.640

Lohman, M. C. (2004). The development of a multirater instrument for assessing employee problem-solving skill. *Human Resource Development Quarterly*, *15*(3).

Lumley, T., & Mcnamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71. https://doi.org/10.1177/026553229501200104

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26*(1), 075–100. https://doi.org/10.1177/0265532208097337

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Mohd Yusri Ibrahim, Mohd Faiz Mohd Yaakob, & Mat Rahimi Yusof. (2019). Communication skills: Top priority of teaching competency. *International Journal of Learning, Teaching and Educational Research*, *18*(8), 17–30. https://doi.org/10.26803/ijlter.18.8.2

Newton, P. E. (2009). The reliability of results from national curriculum testing in England. *Educational Research*, *51*(2), 181–212. https://doi.org/10.1080/00131880902891404

Noor Lide Abu Kassim. (2011). Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, *11*(3), 179–197.

Nor Mashitah, Mariani, Jain Chee, Mohamad Ilmee, Hafiza, & Rosmah. (2015). Penggunaan model pengukuran Rasch many-facet (MFRM) dalam penilaian perkembanagn kanak-kanak berasaskan prestasi. *Jurnal Pendidikan Awal Kanak-Kanak*, *4*, 1–21.

Nor Mashitah Mohd Radzi. (2017). *Pembinaan dan pengesahan instrumen pentaksiran prestasi standard awal pembelajaran dan perkembangan awal kanak-kanak*. Universiti Malaya.

Nur 'Ashiqin Najmuddin. (2011). *Instrumen kemahiran generik pelajar pra-universiti berdasarkan penilaian oleh pensyarah*. Universiti Kebangsaan Malaysia.

Nurul Farahin Ab Aziz, & Siti Mistima Maat. (2021). Kesediaan dan efikasi guru matematik sekolah rendah dalam pengintegrasian teknologi semasa pandemik COVID-19. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, *6*(8), 93–108. https://doi.org/10.47405/mjssh.v6i8.949

OECD. (2013). Preparing teachers for the 21st century: Using evaluation to improve teaching. In *OECD Publishing*. OECD Publishing.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Reseacrh in Nyrsing & Health*, *29*, 489–497. https://doi.org/10.1038/s41590-018-0072-8

Rural, J. D. (2021). Competency in assessment of selected DepEd teachers in National Capital Region. *European Online Journal of Natural and Social Sciences*, *10*(4), 639–646. http://www.european-science.com

Sahin, M. G., Teker, G. T., & Güler, N. (2016). An analysis of peer assessment through many facet Rasch model. *Journal of Education and Practice*, *7*(32), 172–181.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465–493. https://doi.org/10.1177/0265532208094273

Schmidt, F. L., Oh, I.-S., & Shaffer, J. A. (2016). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. In *Validity and Uitility of Selection Methods*. https://doi.org/10.1037/0033-2909.124.2.262

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956–970. https://doi.org/10.1037/0021-9010.85.6.956

Seifert, T., & Feliks, O. (2019). Online self-assessment and peer-assessment as a tool to enhance student-teachers' assessment skills. *Assessment & Evaluation in Higher Education*, *44*(2), 169-185. https://doi.org/10.1080/02602938.2018.1487023

Shin, Y. (2010). A Facets analysis of rater characteristics and rater bias in measuring L2 writing performance. *English Language & Literature Teaching*, *16*(1), 123–142.

Siti Rahayah Ariffin. (2008). *Inovasi dalam pengukuran dan penilaian*. Bangi: Fakulti Pendidikan, Universiti Kebangsaan Malaysia.

Spencer, L. M., & Spencer, S. M. (1993). *Competence at work: Models for superior performance*. United States of America: John Wiley & Sons, Inc.

Springer, D. G., & Bradley, K. D. (2018). Investigating adjudicator bias in concert band evaluations: An application of the many-facets Rasch model. *Musicae Scientiae*, *22*(3), 377–393. https://doi.org/10.1177/1029864917697782

Styck, K. M., Anthony, C. J., Sandilos, L. E., & DiPerna, J. C. (2020). Examining rater effects on the classroom assessment scoring system. *Child Development*, *00*(0), 1–18.

Sumintono, B. (2016). Aplikasi pemodelan Rasch pada asesmen pendidikan: Implementasi penilaian formatif (assessment for learning). *Jurusan Statistika, Institut Teknologi*.

Sunjaya, D. K., Herawati, D., Puteri, D. P., & Sumintono, B. (2020). Development and sensory test of eel cookies for pregnant women with chronic energy deficiency using many facet Rasch model: a preliminary study. *Progress in Nutrition*, *22*(3), 1–11. https://doi.org/10.23751/pn.v22i3.10040

Tomasevic, B. I., Trivic, D. D., Milanovic, V. D., & Ralevic, L. R. (2021). The programme for professional development of chemistry teachers' assessment competency. *Journal of the Serbian Chemical Society*, *86*(10), 997–1010. https://doi.org/10.2298/JSC210710052T

Wang, P., Coetzee, K., Strachan, A., Monteiro, S., & Cheng, L. (2021). Examining rater performance on the CELBAN speaking : A many-facets Rasch measurement analysis. *Canadian Journal of Applied Linguistics*, *23*(2), 73–95.

Warrens, M. J. (2010). A formal proof of a paradox associated with Cohen's kappa. *Journal of Classification*, *27*(3), 322–332. https://doi.org/10.1007/s00357-010-9060-x

Webb, N. M., Shavelson, R. J., & Steedle, J. T. (2018). Generalizability theory in assessment contexts. In *Handbook on measurement, assessment, and evaluation in higher education* (pp. 284–305). https://doi.org/10.4324/9780203142189

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305–319.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA PRESS.

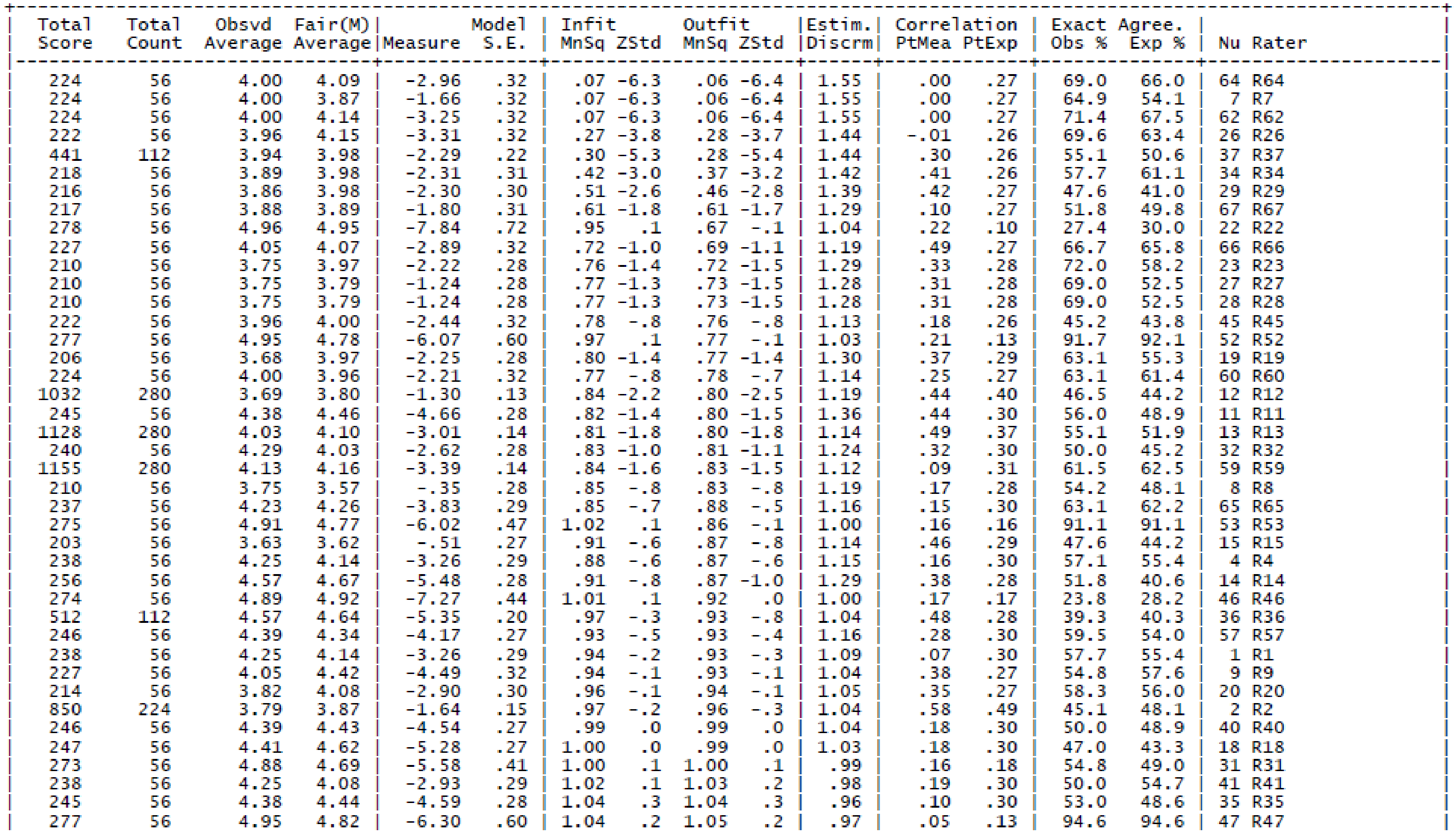
Zhu, W., Ennis, C. D., & Chen, A. (1998). Many-faceted Rasch modeling expert judgment in test development. *Measurement in Physical Education and Exercise Science*, *2*(1), 21–39.

Zhu, Y., Fung, A. S. L., & Yang, L. (2021). A methodologically improved study on raters' personality and rating severity in writing assessment. *SAGE Open*, 1–16.

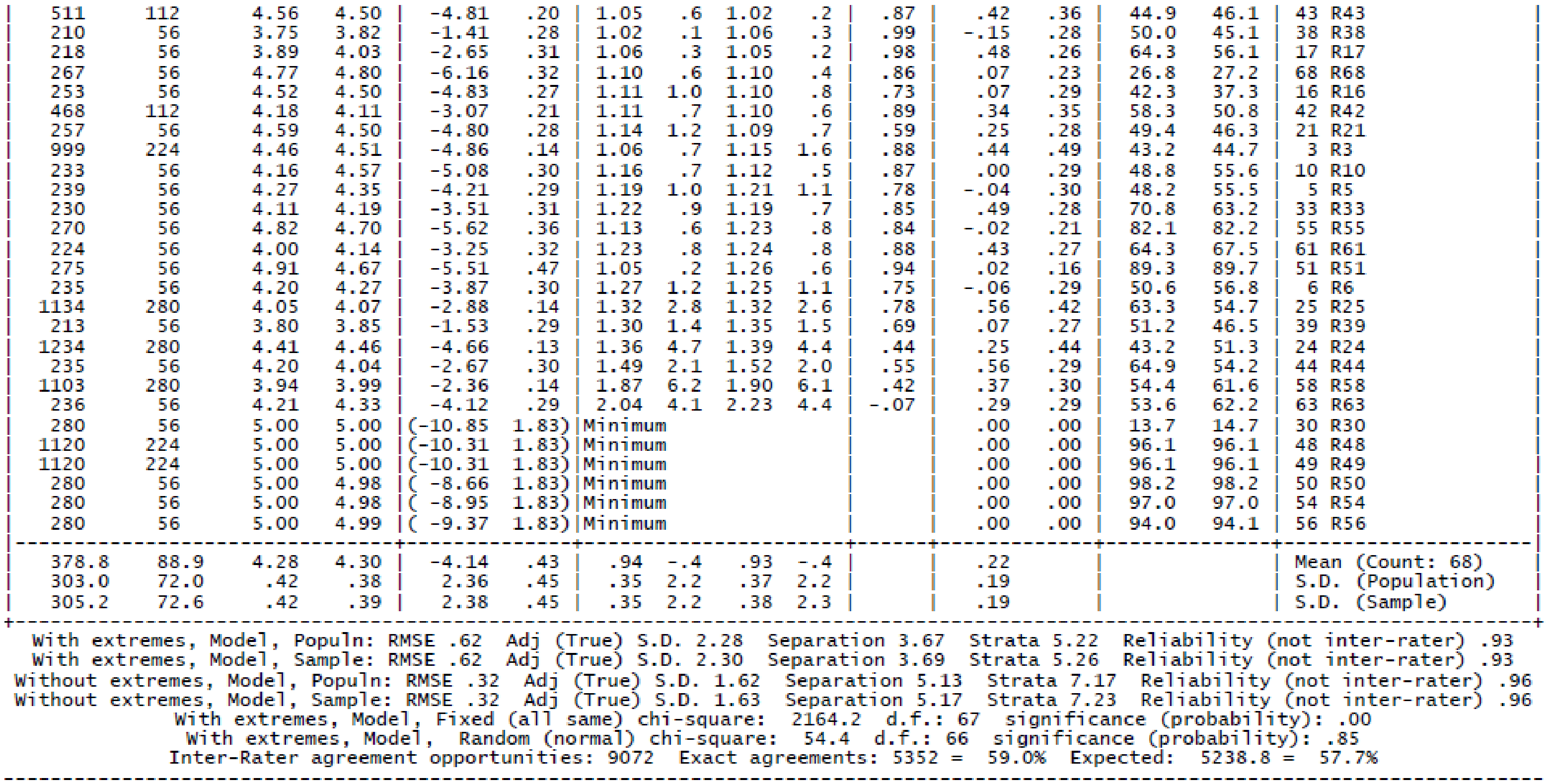
Zuliana Mohd Zabidi, Sumintono, B., & Zuraidah Abdullah. (2021). Enhancing analytic rigor in qualitative analysis : Developing and testing code scheme using many facet Rasch model. *Quality & Quantity*, *55*(2). https://doi.org/10.1007/s11135-021-01152-4

APPENDIX A

Fit Statistics



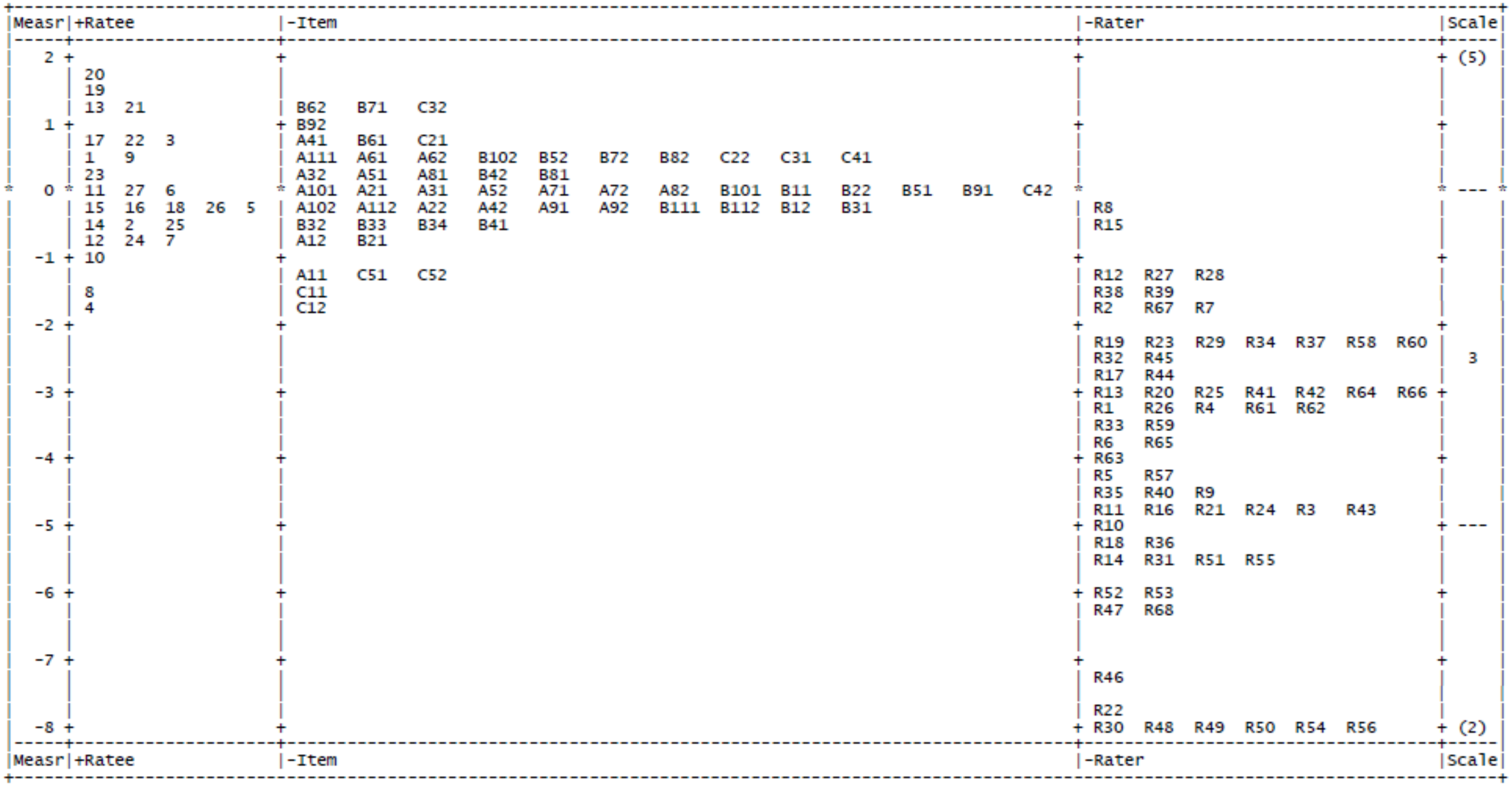
```
+-----------------------------------------------------------------------------------------------------------------------------+
|  Total   Total   Obsvd   Fair(M)|        Model | Infit       Outfit    |Estim.| Correlation | Exact Agree. |                |
|  Score   Count  Average Average|Measure  S.E. | MnSq ZStd  MnSq ZStd |Discrm| PtMea PtExp | Obs %  Exp % | Nu Rater       |
|---------------------------------+--------------+----------------------+------+-------------+--------------+----------------|
|    224      56     4.00    4.09 |  -2.96   .32 |  .07 -6.3   .06 -6.4 | 1.55 |   .00   .27 |  69.0   66.0 | 64 R64         |
|    224      56     4.00    3.87 |  -1.66   .32 |  .07 -6.3   .06 -6.4 | 1.55 |   .00   .27 |  64.9   54.1 |  7 R7          |
|    224      56     4.00    4.14 |  -3.25   .32 |  .07 -6.3   .06 -6.4 | 1.55 |   .00   .27 |  71.4   67.5 | 62 R62         |
|    222      56     3.96    4.15 |  -3.31   .32 |  .27 -3.8   .28 -3.7 | 1.44 |  -.01   .26 |  69.6   63.4 | 26 R26         |
|    441     112     3.94    3.98 |  -2.29   .22 |  .30 -5.3   .28 -5.4 | 1.44 |   .30   .26 |  55.1   50.6 | 37 R37         |
|    218      56     3.89    3.98 |  -2.31   .31 |  .42 -3.0   .37 -3.2 | 1.42 |   .41   .26 |  57.7   61.1 | 34 R34         |
|    216      56     3.86    3.98 |  -2.30   .30 |  .51 -2.6   .46 -2.8 | 1.39 |   .42   .27 |  47.6   41.0 | 29 R29         |
|    217      56     3.88    3.89 |  -1.80   .31 |  .61 -1.8   .61 -1.7 | 1.29 |   .10   .27 |  51.8   49.8 | 67 R67         |
|    278      56     4.96    4.95 |  -7.84   .72 |  .95   .1   .67  -.1 | 1.04 |   .22   .10 |  27.4   30.0 | 22 R22         |
|    227      56     4.05    4.07 |  -2.89   .32 |  .72 -1.0   .69 -1.1 | 1.19 |   .49   .27 |  66.7   65.8 | 66 R66         |
|    210      56     3.75    3.97 |  -2.22   .28 |  .76 -1.4   .72 -1.5 | 1.29 |   .33   .28 |  72.0   58.2 | 23 R23         |
|    210      56     3.75    3.79 |  -1.24   .28 |  .77 -1.3   .73 -1.5 | 1.28 |   .31   .28 |  69.0   52.5 | 27 R27         |
|    210      56     3.75    3.79 |  -1.24   .28 |  .77 -1.3   .73 -1.5 | 1.28 |   .31   .28 |  69.0   52.5 | 28 R28         |
|    222      56     3.96    4.00 |  -2.44   .32 |  .78  -.8   .76  -.8 | 1.13 |   .18   .26 |  45.2   43.8 | 45 R45         |
|    277      56     4.95    4.78 |  -6.07   .60 |  .97   .1   .77  -.1 | 1.03 |   .21   .13 |  91.7   92.1 | 52 R52         |
|    206      56     3.68    3.97 |  -2.25   .28 |  .80 -1.4   .77 -1.4 | 1.30 |   .37   .29 |  63.1   55.3 | 19 R19         |
|    224      56     4.00    3.96 |  -2.21   .32 |  .77  -.8   .78  -.7 | 1.14 |   .25   .27 |  63.1   61.4 | 60 R60         |
|   1032     280     3.69    3.80 |  -1.30   .13 |  .84 -2.2   .80 -2.5 | 1.19 |   .44   .40 |  46.5   44.2 | 12 R12         |
|    245      56     4.38    4.46 |  -4.66   .28 |  .82 -1.4   .80 -1.5 | 1.36 |   .44   .30 |  56.0   48.9 | 11 R11         |
|   1128     280     4.03    4.10 |  -3.01   .14 |  .81 -1.8   .80 -1.8 | 1.14 |   .49   .37 |  55.1   51.9 | 13 R13         |
|    240      56     4.29    4.03 |  -2.62   .28 |  .83 -1.0   .81 -1.1 | 1.24 |   .32   .30 |  50.0   45.2 | 32 R32         |
|   1155     280     4.13    4.16 |  -3.39   .14 |  .84 -1.6   .83 -1.5 | 1.12 |   .09   .31 |  61.5   62.5 | 59 R59         |
|    210      56     3.75    3.57 |   -.35   .28 |  .85  -.8   .83  -.8 | 1.19 |   .17   .28 |  54.2   48.1 |  8 R8          |
|    237      56     4.23    4.26 |  -3.83   .29 |  .85  -.7   .88  -.5 | 1.16 |   .15   .30 |  63.1   62.2 | 65 R65         |
|    275      56     4.91    4.77 |  -6.02   .47 | 1.02   .1   .86  -.1 | 1.00 |   .16   .16 |  91.1   91.1 | 53 R53         |
|    203      56     3.63    3.62 |   -.51   .27 |  .91  -.6   .87  -.8 | 1.14 |   .46   .29 |  47.6   44.2 | 15 R15         |
|    238      56     4.25    4.14 |  -3.26   .29 |  .88  -.6   .87  -.6 | 1.15 |   .16   .30 |  57.1   55.4 |  4 R4          |
|    256      56     4.57    4.67 |  -5.48   .28 |  .91  -.8   .87 -1.0 | 1.29 |   .38   .28 |  51.8   40.6 | 14 R14         |
|    274      56     4.89    4.92 |  -7.27   .44 | 1.01   .1   .92   .0 | 1.00 |   .17   .17 |  23.8   28.2 | 46 R46         |
|    512     112     4.57    4.64 |  -5.35   .20 |  .97  -.3   .93  -.8 | 1.04 |   .48   .28 |  39.3   40.3 | 36 R36         |
|    246      56     4.39    4.34 |  -4.17   .27 |  .93  -.5   .93  -.4 | 1.16 |   .28   .30 |  59.5   54.0 | 57 R57         |
|    238      56     4.25    4.14 |  -3.26   .29 |  .94  -.2   .93  -.3 | 1.09 |   .07   .30 |  57.7   55.4 |  1 R1          |
|    227      56     4.05    4.42 |  -4.49   .32 |  .94  -.1   .93  -.1 | 1.04 |   .38   .27 |  54.8   57.6 |  9 R9          |
|    214      56     3.82    4.08 |  -2.90   .30 |  .96  -.1   .94  -.1 | 1.05 |   .35   .27 |  58.3   56.0 | 20 R20         |
|    850     224     3.79    3.87 |  -1.64   .15 |  .97  -.2   .96  -.3 | 1.04 |   .58   .49 |  45.1   48.1 |  2 R2          |
|    246      56     4.39    4.43 |  -4.54   .27 |  .99   .0   .99   .0 | 1.04 |   .18   .30 |  50.0   48.9 | 40 R40         |
|    247      56     4.41    4.62 |  -5.28   .27 | 1.00   .0   .99   .0 | 1.03 |   .18   .30 |  47.0   43.3 | 18 R18         |
|    273      56     4.88    4.69 |  -5.58   .41 | 1.00   .1  1.00   .1 |  .99 |   .16   .18 |  54.8   49.0 | 31 R31         |
|    238      56     4.25    4.08 |  -2.93   .29 | 1.02   .1  1.03   .2 |  .98 |   .19   .30 |  50.0   54.7 | 41 R41         |
|    245      56     4.38    4.44 |  -4.59   .28 | 1.04   .3  1.04   .3 |  .96 |   .10   .30 |  53.0   48.6 | 35 R35         |
|    277      56     4.95    4.82 |  -6.30   .60 | 1.04   .2  1.05   .2 |  .97 |   .05   .13 |  94.6   94.6 | 47 R47         |
```



```
|   511    112      4.56   4.50 |  -4.81   .20 | 1.05   .6  1.02   .2 |   .87 |   .42    .36 |  44.9   46.1 | 43 R43               |
|   210     56      3.75   3.82 |  -1.41   .28 | 1.02   .1  1.06   .3 |   .99 |  -.15    .28 |  50.0   45.1 | 38 R38               |
|   218     56      3.89   4.03 |  -2.65   .31 | 1.06   .3  1.05   .2 |   .98 |   .48    .26 |  64.3   56.1 | 17 R17               |
|   267     56      4.77   4.80 |  -6.16   .32 | 1.10   .6  1.10   .4 |   .86 |   .07    .23 |  26.8   27.2 | 68 R68               |
|   253     56      4.52   4.50 |  -4.83   .27 | 1.11  1.0  1.10   .8 |   .73 |   .07    .29 |  42.3   37.3 | 16 R16               |
|   468    112      4.18   4.11 |  -3.07   .21 | 1.11   .7  1.10   .6 |   .89 |   .34    .35 |  58.3   50.8 | 42 R42               |
|   257     56      4.59   4.50 |  -4.80   .28 | 1.14  1.2  1.09   .7 |   .59 |   .25    .28 |  49.4   46.3 | 21 R21               |
|   999    224      4.46   4.51 |  -4.86   .14 | 1.06   .7  1.15  1.6 |   .88 |   .44    .49 |  43.2   44.7 |  3 R3                |
|   233     56      4.16   4.57 |  -5.08   .30 | 1.16   .7  1.12   .5 |   .87 |   .00    .29 |  48.8   55.6 | 10 R10               |
|   239     56      4.27   4.35 |  -4.21   .29 | 1.19  1.0  1.21  1.1 |   .78 |  -.04    .30 |  48.2   55.5 |  5 R5                |
|   230     56      4.11   4.19 |  -3.51   .31 | 1.22   .9  1.19   .7 |   .85 |   .49    .28 |  70.8   63.2 | 33 R33               |
|   270     56      4.82   4.70 |  -5.62   .36 | 1.13   .6  1.23   .8 |   .84 |  -.02    .21 |  82.1   82.2 | 55 R55               |
|   224     56      4.00   4.14 |  -3.25   .32 | 1.23   .8  1.24   .8 |   .88 |   .43    .27 |  64.3   67.5 | 61 R61               |
|   275     56      4.91   4.67 |  -5.51   .47 | 1.05   .2  1.26   .6 |   .94 |   .02    .16 |  89.3   89.7 | 51 R51               |
|   235     56      4.20   4.27 |  -3.87   .30 | 1.27  1.2  1.25  1.1 |   .75 |  -.06    .29 |  50.6   56.8 |  6 R6                |
|  1134    280      4.05   4.07 |  -2.88   .14 | 1.32  2.8  1.32  2.6 |   .78 |   .56    .42 |  63.3   54.7 | 25 R25               |
|   213     56      3.80   3.85 |  -1.53   .29 | 1.30  1.4  1.35  1.5 |   .69 |   .07    .27 |  51.2   46.5 | 39 R39               |
|  1234    280      4.41   4.46 |  -4.66   .13 | 1.36  4.7  1.39  4.4 |   .44 |   .25    .44 |  43.2   51.3 | 24 R24               |
|   235     56      4.20   4.04 |  -2.67   .30 | 1.49  2.1  1.52  2.0 |   .55 |   .56    .29 |  64.9   54.2 | 44 R44               |
|  1103    280      3.94   3.99 |  -2.36   .14 | 1.87  6.2  1.90  6.1 |   .42 |   .37    .30 |  54.4   61.6 | 58 R58               |
|   236     56      4.21   4.33 |  -4.12   .29 | 2.04  4.1  2.23  4.4 |  -.07 |   .29    .29 |  53.6   62.2 | 63 R63               |
|   280     56      5.00   5.00 |(-10.85  1.83)|Minimum               |       |   .00    .00 |  13.7   14.7 | 30 R30               |
|  1120    224      5.00   5.00 |(-10.31  1.83)|Minimum               |       |   .00    .00 |  96.1   96.1 | 48 R48               |
|  1120    224      5.00   5.00 |(-10.31  1.83)|Minimum               |       |   .00    .00 |  96.1   96.1 | 49 R49               |
|   280     56      5.00   4.98 |( -8.66  1.83)|Minimum               |       |   .00    .00 |  98.2   98.2 | 50 R50               |
|   280     56      5.00   4.98 |( -8.95  1.83)|Minimum               |       |   .00    .00 |  97.0   97.0 | 54 R54               |
|   280     56      5.00   4.99 |( -9.37  1.83)|Minimum               |       |   .00    .00 |  94.0   94.1 | 56 R56               |
|-------------------------------+--------------+----------------------+-------+--------------+--------------+----------------------|
| 378.8   88.9      4.28   4.30 |  -4.14   .43 |  .94  -.4   .93  -.4 |       |   .22        |              | Mean (Count: 68)     |
| 303.0   72.0       .42    .38 |   2.36   .45 |  .35  2.2   .37  2.2 |       |   .19        |              | S.D. (Population)    |
| 305.2   72.6       .42    .39 |   2.38   .45 |  .35  2.2   .38  2.3 |       |   .19        |              | S.D. (Sample)        |
+----------------------------------------------------------------------------------------------------------------------------------+
  With extremes, Model, Populn: RMSE .62  Adj (True) S.D. 2.28  Separation 3.67  Strata 5.22  Reliability (not inter-rater) .93
  With extremes, Model, Sample: RMSE .62  Adj (True) S.D. 2.30  Separation 3.69  Strata 5.26  Reliability (not inter-rater) .93
 Without extremes, Model, Populn: RMSE .32  Adj (True) S.D. 1.62  Separation 5.13  Strata 7.17  Reliability (not inter-rater) .96
 Without extremes, Model, Sample: RMSE .32  Adj (True) S.D. 1.63  Separation 5.17  Strata 7.23  Reliability (not inter-rater) .96
            With extremes, Model, Fixed (all same) chi-square:  2164.2  d.f.: 67  significance (probability): .00
             With extremes, Model,  Random (normal) chi-square:  54.4  d.f.: 66  significance (probability): .85
            Inter-Rater agreement opportunities: 9072  Exact agreements: 5352 =  59.0%  Expected:  5238.8 =  57.7%
------------------------------------------------------------------------------------------------------------------------------------
```

APPENDIX B

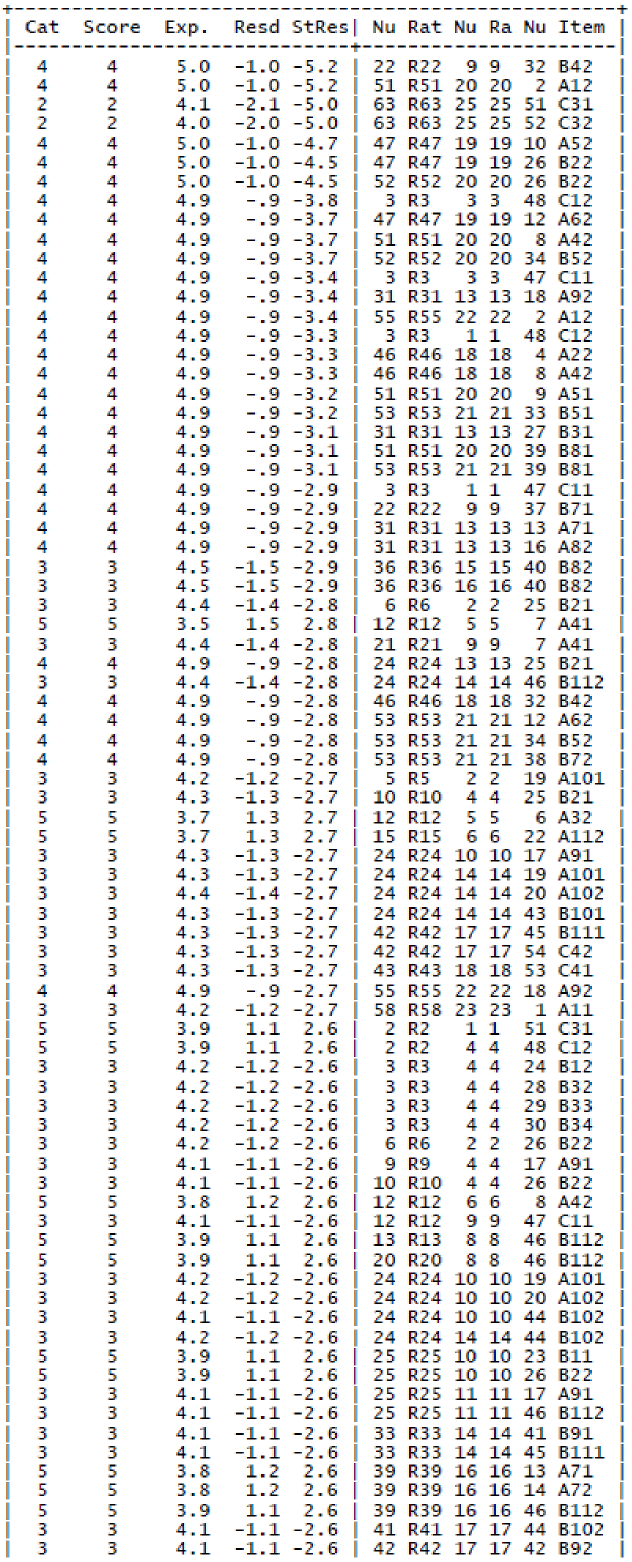
Wright Map



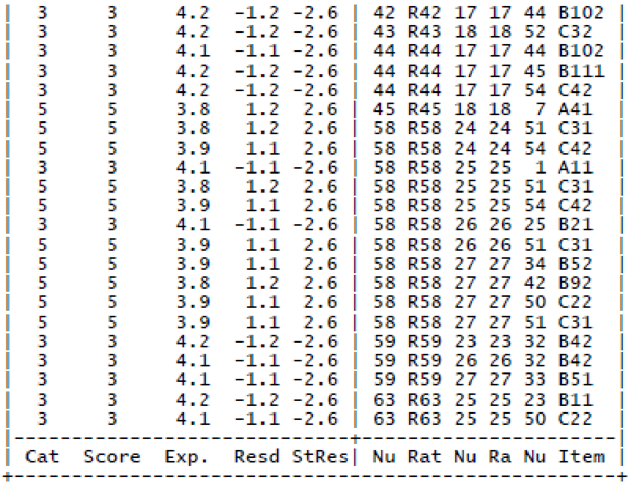
```
+---------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|+Ratee              |-Item                                                                         |-Rater                             |Scale|
|-----+--------------------+------------------------------------------------------------------------------+-----------------------------------+-----|
|   2 +                    +                                                                              +                                   + (5) |
|     | 20                 |                                                                              |                                   |     |
|     | 19                 |                                                                              |                                   |     |
|     | 13  21             | B62   B71   C32                                                              |                                   |     |
|   1 +                    + B92                                                                          +                                   +     |
|     | 17  22  3          | A41   B61   C21                                                              |                                   |     |
|     | 1   9              | A111  A61   A62   B102  B52   B72   B82   C22   C31   C41                    |                                   |     |
|     | 23                 | A32   A51   A81   B42   B81                                                  |                                   |     |
*   0 * 11  27  6          * A101  A21   A31   A52   A71   A72   A82   B101  B11   B22   B51   B91   C42  *                                   * --- *
|     | 15  16  18  26  5  | A102  A112  A22   A42   A91   A92   B111  B112  B12   B31                    | R8                                |     |
|     | 14  2   25         | B32   B33   B34   B41                                                        | R15                               |     |
|     | 12  24  7          | A12   B21                                                                    |                                   |     |
|  -1 + 10                 +                                                                              +                                   +     |
|     |                    | A11   C51   C52                                                              | R12  R27  R28                     |     |
|     | 8                  | C11                                                                          | R38  R39                          |     |
|     | 4                  | C12                                                                          | R2   R67  R7                      |     |
|  -2 +                    +                                                                              +                                   +     |
|     |                    |                                                                              | R19  R23  R29  R34  R37  R58  R60 |     |
|     |                    |                                                                              | R32  R45                          |  3  |
|     |                    |                                                                              | R17  R44                          |     |
|  -3 +                    +                                                                              + R13  R20  R25  R41  R42  R64  R66 +     |
|     |                    |                                                                              | R1   R26  R4   R61  R62           |     |
|     |                    |                                                                              | R33  R59                          |     |
|     |                    |                                                                              | R6   R65                          |     |
|  -4 +                    +                                                                              + R63                               +     |
|     |                    |                                                                              | R5   R57                          |     |
|     |                    |                                                                              | R35  R40  R9                      |     |
|     |                    |                                                                              | R11  R16  R21  R24  R3   R43      |     |
|  -5 +                    +                                                                              + R10                               + --- |
|     |                    |                                                                              | R18  R36                          |     |
|     |                    |                                                                              | R14  R31  R51  R55                |     |
|     |                    |                                                                              |                                   |     |
|  -6 +                    +                                                                              + R52  R53                          +     |
|     |                    |                                                                              | R47  R68                          |     |
|     |                    |                                                                              |                                   |     |
|     |                    |                                                                              |                                   |     |
|  -7 +                    +                                                                              +                                   +     |
|     |                    |                                                                              | R46                               |     |
|     |                    |                                                                              |                                   |     |
|     |                    |                                                                              | R22                               |     |
|  -8 +                    +                                                                              + R30  R48  R49  R50  R54  R56      + (2) |
|-----+--------------------+------------------------------------------------------------------------------+-----------------------------------+-----|
|Measr|+Ratee              |-Item                                                                         |-Rater                             |Scale|
+---------------------------------------------------------------------------------------------------------------------------------------------------+
```

APPENDIX C

Unexpected Response



```
+-----------------------------------------------------+
| Cat  Score  Exp.  Resd StRes| Nu Rat Nu Ra Nu Item  |
|-----------------------------+-----------------------|
|  4     4     5.0  -1.0 -5.2 | 22 R22  9 9  32 B42   |
|  4     4     5.0  -1.0 -5.2 | 51 R51 20 20  2 A12   |
|  2     2     4.1  -2.1 -5.0 | 63 R63 25 25 51 C31   |
|  2     2     4.0  -2.0 -5.0 | 63 R63 25 25 52 C32   |
|  4     4     5.0  -1.0 -4.7 | 47 R47 19 19 10 A52   |
|  4     4     5.0  -1.0 -4.5 | 47 R47 19 19 26 B22   |
|  4     4     5.0  -1.0 -4.5 | 52 R52 20 20 26 B22   |
|  4     4     4.9   -.9 -3.8 |  3 R3   3 3  48 C12   |
|  4     4     4.9   -.9 -3.7 | 47 R47 19 19 12 A62   |
|  4     4     4.9   -.9 -3.7 | 51 R51 20 20  8 A42   |
|  4     4     4.9   -.9 -3.7 | 52 R52 20 20 34 B52   |
|  4     4     4.9   -.9 -3.4 |  3 R3   3 3  47 C11   |
|  4     4     4.9   -.9 -3.4 | 31 R31 13 13 18 A92   |
|  4     4     4.9   -.9 -3.4 | 55 R55 22 22  2 A12   |
|  4     4     4.9   -.9 -3.3 |  3 R3   1 1  48 C12   |
|  4     4     4.9   -.9 -3.3 | 46 R46 18 18  4 A22   |
|  4     4     4.9   -.9 -3.3 | 46 R46 18 18  8 A42   |
|  4     4     4.9   -.9 -3.2 | 51 R51 20 20  9 A51   |
|  4     4     4.9   -.9 -3.2 | 53 R53 21 21 33 B51   |
|  4     4     4.9   -.9 -3.1 | 31 R31 13 13 27 B31   |
|  4     4     4.9   -.9 -3.1 | 51 R51 20 20 39 B81   |
|  4     4     4.9   -.9 -3.1 | 53 R53 21 21 39 B81   |
|  4     4     4.9   -.9 -2.9 |  3 R3   1 1  47 C11   |
|  4     4     4.9   -.9 -2.9 | 22 R22  9 9  37 B71   |
|  4     4     4.9   -.9 -2.9 | 31 R31 13 13 13 A71   |
|  4     4     4.9   -.9 -2.9 | 31 R31 13 13 16 A82   |
|  3     3     4.5  -1.5 -2.9 | 36 R36 15 15 40 B82   |
|  3     3     4.5  -1.5 -2.9 | 36 R36 16 16 40 B82   |
|  3     3     4.4  -1.4 -2.8 |  6 R6   2 2  25 B21   |
|  5     5     3.5   1.5  2.8 | 12 R12  5 5   7 A41   |
|  3     3     4.4  -1.4 -2.8 | 21 R21  9 9   7 A41   |
|  4     4     4.9   -.9 -2.8 | 24 R24 13 13 25 B21   |
|  3     3     4.4  -1.4 -2.8 | 24 R24 14 14 46 B112  |
|  4     4     4.9   -.9 -2.8 | 46 R46 18 18 32 B42   |
|  4     4     4.9   -.9 -2.8 | 53 R53 21 21 12 A62   |
|  4     4     4.9   -.9 -2.8 | 53 R53 21 21 34 B52   |
|  4     4     4.9   -.9 -2.8 | 53 R53 21 21 38 B72   |
|  3     3     4.2  -1.2 -2.7 |  5 R5   2 2  19 A101  |
|  3     3     4.3  -1.3 -2.7 | 10 R10  4 4  25 B21   |
|  5     5     3.7   1.3  2.7 | 12 R12  5 5   6 A32   |
|  5     5     3.7   1.3  2.7 | 15 R15  6 6  22 A112  |
|  3     3     4.3  -1.3 -2.7 | 24 R24 10 10 17 A91   |
|  3     3     4.3  -1.3 -2.7 | 24 R24 14 14 19 A101  |
|  3     3     4.4  -1.4 -2.7 | 24 R24 14 14 20 A102  |
|  3     3     4.3  -1.3 -2.7 | 24 R24 14 14 43 B101  |
|  3     3     4.3  -1.3 -2.7 | 42 R42 17 17 45 B111  |
|  3     3     4.3  -1.3 -2.7 | 42 R42 17 17 54 C42   |
|  3     3     4.3  -1.3 -2.7 | 43 R43 18 18 53 C41   |
|  4     4     4.9   -.9 -2.7 | 55 R55 22 22 18 A92   |
|  3     3     4.2  -1.2 -2.7 | 58 R58 23 23  1 A11   |
|  5     5     3.9   1.1  2.6 |  2 R2   1 1  51 C31   |
|  5     5     3.9   1.1  2.6 |  2 R2   4 4  48 C12   |
|  3     3     4.2  -1.2 -2.6 |  3 R3   4 4  24 B12   |
|  3     3     4.2  -1.2 -2.6 |  3 R3   4 4  28 B32   |
|  3     3     4.2  -1.2 -2.6 |  3 R3   4 4  29 B33   |
|  3     3     4.2  -1.2 -2.6 |  3 R3   4 4  30 B34   |
|  3     3     4.2  -1.2 -2.6 |  6 R6   2 2  26 B22   |
|  3     3     4.1  -1.1 -2.6 |  9 R9   4 4  17 A91   |
|  3     3     4.1  -1.1 -2.6 | 10 R10  4 4  26 B22   |
|  5     5     3.8   1.2  2.6 | 12 R12  6 6   8 A42   |
|  3     3     4.1  -1.1 -2.6 | 12 R12  9 9  47 C11   |
|  5     5     3.9   1.1  2.6 | 13 R13  8 8  46 B112  |
|  5     5     3.9   1.1  2.6 | 20 R20  8 8  46 B112  |
|  3     3     4.2  -1.2 -2.6 | 24 R24 10 10 19 A101  |
|  3     3     4.2  -1.2 -2.6 | 24 R24 10 10 20 A102  |
|  3     3     4.1  -1.1 -2.6 | 24 R24 10 10 44 B102  |
|  3     3     4.2  -1.2 -2.6 | 24 R24 14 14 44 B102  |
|  5     5     3.9   1.1  2.6 | 25 R25 10 10 23 B11   |
|  5     5     3.9   1.1  2.6 | 25 R25 10 10 26 B22   |
|  3     3     4.1  -1.1 -2.6 | 25 R25 11 11 17 A91   |
|  3     3     4.1  -1.1 -2.6 | 25 R25 11 11 46 B112  |
|  3     3     4.1  -1.1 -2.6 | 33 R33 14 14 41 B91   |
|  3     3     4.1  -1.1 -2.6 | 33 R33 14 14 45 B111  |
|  5     5     3.8   1.2  2.6 | 39 R39 16 16 13 A71   |
|  5     5     3.8   1.2  2.6 | 39 R39 16 16 14 A72   |
|  5     5     3.9   1.1  2.6 | 39 R39 16 16 46 B112  |
|  3     3     4.1  -1.1 -2.6 | 41 R41 17 17 44 B102  |
|  3     3     4.1  -1.1 -2.6 | 42 R42 17 17 42 B92   |
```



```
|  3     3     4.2  -1.2 -2.6 | 42 R42 17 17 44 B102 |
|  3     3     4.2  -1.2 -2.6 | 43 R43 18 18 52 C32  |
|  3     3     4.1  -1.1 -2.6 | 44 R44 17 17 44 B102 |
|  3     3     4.2  -1.2 -2.6 | 44 R44 17 17 45 B111 |
|  3     3     4.2  -1.2 -2.6 | 44 R44 17 17 54 C42  |
|  5     5     3.8   1.2  2.6 | 45 R45 18 18  7 A41  |
|  5     5     3.8   1.2  2.6 | 58 R58 24 24 51 C31  |
|  5     5     3.9   1.1  2.6 | 58 R58 24 24 54 C42  |
|  3     3     4.1  -1.1 -2.6 | 58 R58 25 25  1 A11  |
|  5     5     3.8   1.2  2.6 | 58 R58 25 25 51 C31  |
|  5     5     3.9   1.1  2.6 | 58 R58 25 25 54 C42  |
|  3     3     4.1  -1.1 -2.6 | 58 R58 26 26 25 B21  |
|  5     5     3.9   1.1  2.6 | 58 R58 26 26 51 C31  |
|  5     5     3.9   1.1  2.6 | 58 R58 27 27 34 B52  |
|  5     5     3.8   1.2  2.6 | 58 R58 27 27 42 B92  |
|  5     5     3.9   1.1  2.6 | 58 R58 27 27 50 C22  |
|  5     5     3.9   1.1  2.6 | 58 R58 27 27 51 C31  |
|  3     3     4.2  -1.2 -2.6 | 59 R59 23 23 32 B42  |
|  3     3     4.1  -1.1 -2.6 | 59 R59 26 26 32 B42  |
|  3     3     4.1  -1.1 -2.6 | 59 R59 27 27 33 B51  |
|  3     3     4.2  -1.2 -2.6 | 63 R63 25 25 23 B11  |
|  3     3     4.1  -1.1 -2.6 | 63 R63 25 25 50 C22  |
|-----------------------------+----------------------|
| Cat   Score  Exp.  Resd StRes| Nu Rat Nu Ra Nu Item |
+----------------------------------------------------+
```