RATEE COMPETENCY LEVEL IN CLASSROOM ASSESSMENT: APPLICATION OF MANY FACET RASCH MODEL

***Rosyafinaz Mohamat[1]**
Bambang Sumintono[2]
Harris Shah Abdul Hamid[3]

[1] Curriculum Development Unit, Ministry of Education Malaysia
[2] Faculty of Education, Universitas Islam Internasional Indonesia, Indonesia
[3] Faculty of Management, Education & Humanities, University College MAIWP International

**rosyafinaz@moe.gov.my*

Abstract: This study examines teachers' Classroom Assessment (CA) competency using the Many Facet Rasch Model (MFRM) analysis. Instruments with good psychometric characteristics can guarantee more accurate and fair measurement to respondents. If such instruments are not developed, it is feared that teachers cannot identify their competency level in the CA. The instrument consists of 56 items built based on three primary 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 262 raters: The Head of Mathematics and Science Department, The Head of Mathematics Panel, and the Mathematics Teacher to assess 100 ratees. The ratees involved in this study are 100 secondary school Mathematics teachers from Selangor. The results show that among the advantages of MFRM are that it can determine the ability and consistency level of the ratees and also detect unexpected responses by the ratees. This study indicates that MFRM is an alternative model suitable to overcome the limitations in Classical Test Theory (CTT) and statistical models in multi-rater analysis. MFRM has the advantage of providing complete information and contributes to understanding the consistent analysis of the ratees' ability with quantitative evidence support. Furthermore, MFRM can produce better and more precise measurements and make it easier for researchers to communicate the findings.

Keywords: Many Facet Rasch Model, Competency, Classroom Assessment, Ratee Ability, Multi-Rater Analysis

INTRODUCTION

Teacher competence issues have attracted many researchers and the attention of many stakeholders in Malaysia to how teachers educate students (Muhd Khaizer et al., 2020). The basic concept of competence is that the individual's ability matches the assigned task and can optimise individual performance (Boyatzis, 2008). The most common analysis method is Classical Test Theory (CTT), which is optimal when only one rater evaluates all the ratees (Nur 'Ashiqin, 2011). The reliability of CTT will improve if the raters agree on their judgments more consistently (Noor Lide, 2011). The reliability and validity of the performance assessment can be increased, and conclusions about the ratee's ability can be more accurate due to the Many Facet Rasch Model (MFRM) (Engelhard, 1994).

MFRM has many benefits that help it overcome the limitations of the CTT approach. Because self-assessment is subjective, an individual's response is probably lower or higher than their ability and does not accurately reflect their behaviour (bias). The bias might occur because self-assessment relies on respondents' honesty and does not always represent actual behaviour. This research used a multi-rater approach that does not rely on self-assessment. As a result, the article aims to demonstrate how the MFRM can help measure the teachers' competency level more accurately and precisely.

LITERATURE REVIEW

How to Measure the Teachers' Competency in Classroom Assessment

The subjective assessment of the rater can affect the reliability and validity of the ratee's performance (Schaefer, 2008). Using a single rater can lead to a biased judgment (Matsuno, 2009). Self-assessment and peer-assessment have grown in education because they can overcome this constraint (Hargreaves et al., 2002). The multi-rater method is steadier and more accurate than self-assessment and has higher reliability (Calhoun et al., 2011; Goffin

& Jackson, 1992; Lohman, 2004). When more raters are involved in classroom assessment, for example, the reliability of the results is increased (Kane & Staiger, 2012).

The previous related study suggests that a researcher can obtain a more accurate and fair assessment with a multi-rater method. The multi-rater practice, which includes peer assessment, self-assessment, and assessment by superiors or subordinates, has grown in popularity in determining an individual's job performance (Scullen et al., 2000). It is advised that more than one rater be included when assessing teacher quality through performance assessment, as the inclusion of numerous raters is often viewed as the "secret" to effective teacher assessment methods (OECD, 2013).

Teachers' Competency Influenced by Variability among Ratee

Teacher skills in assessment tasks and responsibilities (monitoring, analyzing, communicating, implementing and feedback) can be influenced by the demographic characteristics of the teacher, especially the career level and education assessment. (DeLuca et al., 2018). The teacher's experience of teaching contributes toward developing the teacher's competence in assessing the student's performance (Al-Bahlani, 2019). Another significant variable connected to competency which has been identified and is extensively investigated by researchers is gender (Gunal et al., 2015; Kursad, 2022).

It is critical to determine whether competency differs by gender to take the required precautions if one group has lower competency than the other (Kursad, 2022). The previous study on competency perceptions revealed gender as an ineffectual independent variable. (Kursad, 2022). The study by Gunal et al. (2015) found that female students have a more positive attitude towards measurement and evaluation lessons than males. The study also mentioned that a positive attitude towards measurement and evaluation could improve the assessment methods used.

Common Method Used in Multi-Rater Analysis

Various methods based on the CTT approach have been widely utilised to determine the consistency of raters. The Cohen Kappa technique, for example, assesses consistency between two raters by excluding agreement between them (Hsu & Field, 2003). Next, the Fleiss Kappa approach provides statistical comparison interpretations that are easier to understand than the Cohen Kappa method, making determining the rater's agreement more difficult to interpret (Allen, 2017).

The following method is Generalizability Theory (G theory) by Lee Cronbach, which was developed to examine rater reliability and isolate and assume the various sources (Brennan, 2010; Webb et al., 2018). The G theory is an enhanced statistical theory of CTT that allows for a more accurate assessment of reliability related to behavioural measures by assuming diverse error sources (Nor Mashitah, 2017). Another method for determining the validity of the overall content of an instrument in multi-rater contexts is the Content Validity Index (CVI), which is produced using the average Content Validity Ratio (CVR) (Lindell & Brandt, 1999). By converting ordinal scale data into two categories (relevant or irrelevant), CVI gives direct information about the rater's agreement (Polit & Beck, 2006).

The Problem of Measuring Ratee Ability in Classical Test Theory

The CTT methodology has some disadvantages when it comes to multi-rater analysis approaches. If there are only two raters, Cohen Kappa can be used, while Fleiss Kappa can be used with more than two raters, but only with nominal data categories (Cohen, 1960; Fleiss & Cohen, 1973). The Fleiss Kappa approach, on the other hand, is doubtful since it is based on the assumption of homogeneity and is challenging to apply to polytomous data (Allen, 2017; Bartok & Burzler, 2020; Warrens, 2010). The Fleiss Kappa method is also unable to discern if the raters are guessing throughout the scoring process and unable to determine the severity level of the raters (Allen, 2017).

Furthermore, the CTT-based internal consistency measurement has a weakness in that it cannot systematically distinguish the raters, such as when the severity level of the raters is consistent with all ratees (Newton, 2009). Even though G Theory has some advantages over the frequently used CTT approach, it is highly complex and challenging for the reader to accept and understand the interpretation (Brennan, 2010; Webb et al., 2018). The G Theory also has some weaknesses, such as not determining the severity level of the raters and not including the rater's error in the scale testing explanation (Zhu et al., 1998).

Moreover, the CVI method has several limitations, including involving only two categories of an ordinal scale, the rater's agreement index is likely to decrease as the number of raters increases, determining rater's agreement using the average value approach, and only focusing on item suitability without involving scale analysis to ensure accurate construct measurements (Polit & Beck, 2006). The CVR approach can only assess dichotomous data (Lindell & Brandt, 1999).

MFRM in Research

A study by Maryati (2019) used a multi-rater approach to assess teacher professionalism on pedagogical content knowledge. The study found that MFRM produced clear information on the ability of 20 teachers assessed using six experts, showing that MFRM could produce accurate analysis using small samples. The findings were similar to the study by Nurul Nadia et al. (2018) stated that the multi-rater approach can produce a more accurate assessment and showed the capability of MRFM to place individual responses, items, and evaluations on the same interval scale.

The advantages of MFRM have also been proven in a study by Fahmina et al. (2019), which used MFRM to analyze the Computerized Testlet Instrument to Measure Chemical Literature Capabilities which was assessed by nine raters against 21 items based on five aspects of assessment. The findings showed that MFRM could provide detailed information on the reliability and separation index, rater’s agreement percentage, the most difficult or easily achievable aspects of the item and the arrangement of the rater’s severity level.

A study by Norzetty and Sumintono (2017) examined the influence of the tactics of 18 leaders from six departments in the Ministry of Education. Each leader was assessed by ten individuals that work in the same department as the leader. The data was obtained using the Influence Tactics Behaviour (IBQ) instrument. The MFRM was used to identify the difficulty level of each item and the ability level for each ratee. The findings showed that leaders have three levels of ability: low, medium, and high. The findings also showed the advantages of MFRM that can identify commonly or rarely used tactics by leaders. Furthermore, the study also showed that MFRM could provide detailed information such as the demographic characteristics of the respondents and found that the leader's background did not influence their ability level.

A study by Wu dan Tan (2016) used the MFRM to identify the rater's behaviour and showed how it influenced student performance. The study showed the advantages of MFRM in producing data that allowed researchers to handle practical issues on assessment. Compared to CTT, MFRM has the advantage of detecting incorrect rater responses, inappropriate judgement patterns, and missing data (Fahmina et al., 2019; Goodwin & Leech, 2003). Furthermore, MFRM can provide more detailed information on ratee, rater, and criteria; the analysis procedure is easier and faster; it can detect missing data and consider the difference between the severity of the rater and the difficulty of the criteria measured (Eckes, 2015). This statement clearly shows that MFRM is a viable alternative model for overcoming the weaknesses of CTT statistical models.

METHODOLOGY

Instrumentation

The instrument used in this study measures teachers' competency in the CA. This instrument contains 56 items that are classified into three parts: knowledge in CA (22 items), skills in CA (24 items), and attitude towards CA (10 items). The determination of the constructs is based on an analysis of eight competency models and 13 existing competency instruments, which have been adjusted to Classroom Assessment Implementation Guidelines (Second Edition) by Bahagian Pembangunan Kurikulum (2019). All items were assessed using a 5-point Likert scale; the higher the score, the better the ratee's performance. The raters will respond to all items to evaluate the ratee's ability.

The Respondents

Selangor is the populous state in Malaysia and can be used to describe the country's characteristics. This study's population is Mathematics teachers working in government secondary schools in Selangor. Selangor has the most teachers compared to other states. Apart from that, Selangor, after Johor, is the state with the most secondary schools. The respondents in this study were chosen using a variety of sampling techniques. The cluster sampling technique was used to divide Selangor into ten districts. Then, the basic random sampling technique was used to determine the four districts and six schools for each of the four districts involved. A simple random sampling technique was used to select the five teachers to be evaluated (ratee) for each school. Finally, the five raters for each ratee were identified using the purposive sampling technique.

The respondents involved in this study were 324 raters who evaluated 108 teachers. Each ratee was supposed to be rated by five raters. The five raters consist of self-assessment, The School Improvement Specialist Coaches

(SISC+), The Head of Mathematics & Science Department, The Head of Mathematics Panel, and the Mathematic teachers. But after the researchers completed the data collection, there were 57 teachers rated by five raters, 18 teachers were rated by four raters, 23 teachers were rated by three raters, and ten teachers were rated by three raters. After the data screening process, the respondents involved in this study were 262 raters to evaluate 100 teachers.

Table 1. Background Information of the Respondents

Demographic	Factors	Frequency	Percentage (%)
Gender	Male	28	10.69
	Female	234	89.31
Age	20-29 years	7	2.67
	30-39 years	111	42.37
	40-49 years	107	40.84
	50-60 years	37	14.12
Ethnicity	Malay	224	85.50
	Chinese	17	6.49
	Indian	18	6.78
	Others	3	1.15
Position	SISC+	6	2.29
	The Head of Mathematics & Science Department	17	6.49
	The Head of Mathematics Panel	19	7.25
	Mathematics Teacher	220	83.97
Experience	1-9 years	51	19.47
	10-19 years	155	59.16
	20-29 years	56	21.37
	30-39 years	0	0.00

Measurement Model

The researchers analysed the data with MFRM to evaluate the severity, consistency, and bias interaction in the raters' assessments. Fit statistics are necessary for researchers to determine the accuracy of the data fit to the Rasch model (Siti Rahayah, 2008). The value of Infit MnSq and Outfit MnSq in the fit statistic indicates how consistently the rater did the judgement. According to the Rasch model parameters, MnSq = 1 suggests that the data is ideal. In the fit statistic, a value of MnSq of 0.5 to 1.5 is acceptable (Bond & Fox, 2015).

The data is accepted if the data reliability index is more than 0.65 (Bond & Fox, 2015). The separation index was calculated using the assumptions or estimations of respondents' separation or differences depending on their level of ability on the measured variables (Wright & Masters, 1982). If the separation index is greater than 2, it indicates a good and acceptable result (Linacre, 2006). As an indicator of good unidimensionality, Rasch analysis demands at least 40% of raw variance explained by measures (Bond & Fox, 2015). Meanwhile, to analyse the differences in the competency level between male and female ratees, an independent samples test was conducted using the IBM SPSS Statistics 25 program.

RESULTS

Reliability and Construct Validity

The researchers used the value of the reliability and validity index from the MFRM analysis findings to determine the assessment's reliability.

Table 2. MFRM Analysis Findings

	Rater	Ratee	Item
N	262	100	56
Mean	-3.71	0.00	0.00
Standard Deviation (SD)	3.14	1.03	0.50
Standard Error (SE)	0.39	0.21	0.14
Separation Index	7.46	4.70	3.33
Strata	10.28	6.59	4.78
Reliability Index	0.98	0.96	0.92
Significance (probability) (p)	0.00	0.00	0.00
Observed Exact Agreements (%)		52.1	
Expected Agreements (%)		52.2	
Variance explained by Rasch measures (%)		62.82	

The rater's reliability index is 0.98, which is a good score. The rater separation index of 7.46 is good because it is larger than 3. The significance (probability) score of p = 0.00 indicated that the raters' severity levels differed significantly, and the raters' judgments were very consistent. This data revealed that each panel evaluates at a different severity level. The actual rater agreement percentage was 52.1%, compared to the expected rater agreement percentage of 52.2%. The rater's judgement is good and not homogeneous, indicating that it has excellent inter-rater reliability and meets the Rasch Model's predictions.

The ratee's reliability index is 0.96, which is a good score. The ratee separation index of 4.70 is also good because it is larger than 3. A significant difference in the ratee's ability level was indicated by the p = 0.00 significance (probability) value. As a result, different levels of ratee ability exist. The item's reliability index is 0.92, which is a good value. The item separation index of 3.33 is also good since it is higher than 3. The significance (probability) value is p = 0.00, showing a significant difference in the item's difficulty level. According to these findings, the instrument has a variety of item difficulty levels.

The instrument's unidimensionality was also a determinant in guaranteeing that it could only measure in one direction. The percentage of variance explained by Rasch measures is 62.82% showing that the instruments have a high unidimensionality. The percentage of variance explained by Rasch measures must account for at least 40% (Engelhard & Wind, 2018).

Rating Scale

The researchers conducted the rating scale analysis to verify that the five scales correctly measured the teacher's competency in CA. The rating scale analysis aims to evaluate the scale's validity to analyse the validated data appropriately. Items that respondents easily agree on are assumed to be given a high score in rating scale analysis (Wright & Masters, 1982). The value of the Rasch-Andrich Threshold, which may indicate the threshold value, can be used in the Rasch model to describe the scale for each item (Siti Rahayah, 2008a).

This threshold value can aid researchers in determining an individual's turning point while transitioning from one scale to the next. The threshold value, s, in the range 1.4-5.0, shows that the classification is applicable. If the value of s is less than 1.4, the rating should be collapsed and separated if the value of s is greater than 5.0 (Linacre, 2006). The response structure on the logit scale can be shown using the rating scale analysis. The probability curves depict the responses to the measurement categories in categories 1 to 5. When two neighbouring curves overlap, the scale probabilities for both categories are the same.

The researchers analysed the rating scale to guarantee that the respondents understood and differentiated the five-category scale used. This analysis aims to determine whether scales should be kept, collapsed, or separated depending on Rasch-Andrich threshold values. The criteria to be emphasised in determining the suitability of the scale category used are (1) a minimum of 10 responses for each scale category; (2) the average measure increased along with the scale category; (3) the outfit MnSq value is less than 2; and (4) threshold values in the range 1.4 s 5.0 (Bond & Fox, 2015). The researchers do not need to meet all four criteria in evaluating the scale category suitability in the rating scale analysis since they can refer to the best quality that fits the standards (Linacre, 2002).

Table 3. Structure Calibration of Rating Scale

Category Total	Counts Used	Percentage (%)	Average Measure	Outfit MnSq	Rasch Andrich Threshold
2	126	1	-6.72	0.80	-
3	2611	12	-0.73	0.80	-7.20
4	14093	67	3.79	0.90	0.07
5	4170	20	7.96	0.80	7.13

As seen in the second column, the frequency values for all scale categories are good because there are more than 10. The percentage for each frequency is shown in the third column. According to the findings, none of the respondents used scale 1, just 1% (n=126) used scale 2, 12% (n=2611) used scale 3, 67 % (n=14093) used scale 4, and 20% (n=24170) used scale 5. In comparison to the other scales, scale 4 has the highest frequency. The average measure for each scale category is shown in the fourth column.

Based on the findings, the average measure for scale 2 is -6.72, the average measure for scale 3 is -0.73, the average measure for scale 4 is 3.79, and the average measure for scale 2 is 7.96. According to the findings, the average measure values rise with the scale category from -6.72 to 7.96. The MnSq outfit values for each scale category were acceptable because they were in the range of 0.80 to 0.90, and none surpassed 2.0. The researchers then looked at the MnSq values in the fifth column, which should not exceed 2.0. Meanwhile, the value of the Rasch-Andrich threshold is shown in the sixth column.

Table 4. Rating Scale Analysis

Difference Between Range	(1.4 < s < 5)
1-2	-
2-3	[0 - (- 7.20)] = 7.20
3-4	[- 7.20 – (0.07)] = -7.27
4-5	[0.07 – (7.13)] =7.06

The results showed that the respondents were highly good at distinguishing between different scales. Even though the s value found was unacceptable, the researchers kept all scale categories. Each scale category with a distinct peak significantly increased its threshold value. In addition, the findings meet other criteria, such as the number of frequencies being larger than 10, the average measure value increasing with the scale category, the outfit value being less than 2, and the peaks for each scale being evident and not overlapping. Finally, the findings suggest that the respondents can read, comprehend, and differentiate the scale categories employed in the instrument.

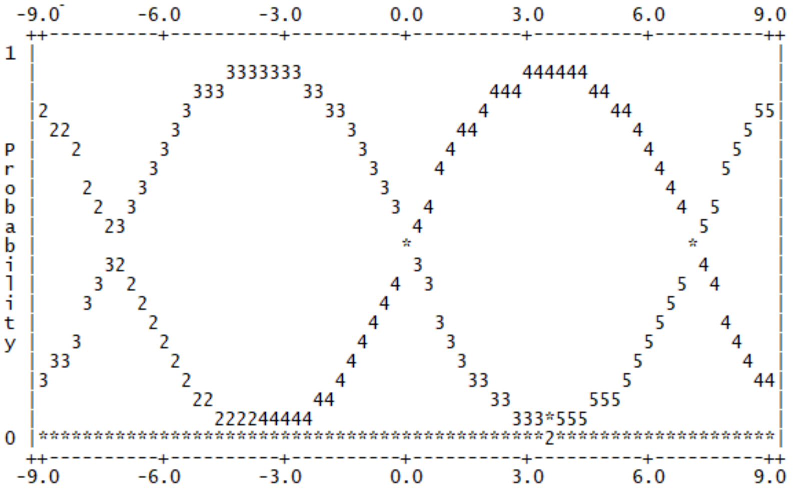


Figure 1. Rating Scale Calibration Structure Analysis

Ratee Logit

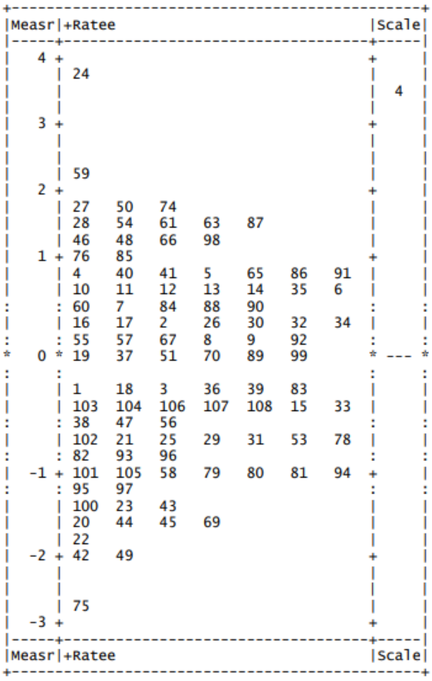
Because items are supposed to have different discrimination indexes, and each item is measured solely by the difficulty parameter, item difficulty, rater severity, and individual ability are placed on the same logit scale when using MFRM (Bond & Fox, 2015). The raters' judgement of the ratee's ability level, the item's difficulty level, and the rater's severity level is represented by the logit measure value. Researchers can use Wright's map to compare individual abilities and item difficulties (Boone, 2020).

The benchmark to determine the teachers' competency level in CA is based on the logit values. The results showed the min logit value for the ratee is 0.00 with a logit standard deviation of 1.03. Therefore, ratees with a logit value of less than 0.00 are classified as having low competency. In contrast, ratees who obtained a logit value of more than 0.00 are classified as highly competent. The findings show that there are not many differences between the number of high-ability and low-ability ratees. There are 53 ratees categorized as high-capable individuals and 47 ratees classified as low-ability individuals.

Table 5. Ratees' Competency Level (N=100)

Ratee Competency Level	Logit Value	Ratee	Frequency	Percentage
Very high	+1.03 and above	24 59 27 50 74 63 87 28 61 54 48 98 66 46 76 85	16	16%
High	0.00 to +1.03	4 40 41 91 65 5 86 13 10 11 12 14 7 60 90 88 84 6 35 2 55 34 17 26 30 32 92 67 16 57 8 9 51 89 19 70 99	37	37%
Moderate	-1.03 to 0.00	37 3 1 83 36 39 18 33 47 106 38 15 104 108 103 107 56 25 78 53 21 93 102 82 31 29 96 95 81 101 94 79 97	33	33%
Low	-1.03 and below	105 80 58 100 23 43 44 69 20 45 22 42 49 75	14	14%

Based on the logit value obtained, the researchers decided to categorize the ratees' competency level in more detail. A small logit measurement value indicates a low individual ability level, while a large one indicates a high individual ability level (Boone et al., 2014). Figure 2 illustrates the four-ability level of ratees. The location of item ratee 24 at the top of the chart indicated that the ratee had the highest competency level. In contrast, ratee 75 below shows that the ratee had the lowest competency level.



```
+-------------------------------------------------------+
|Measr|+Ratee                                     |Scale|
|-----+-------------------------------------------+-----|
|   4 +                                           +     |
|     | 24                                        |     |
|     |                                           |  4  |
|     |                                           |     |
|   3 +                                           +     |
|     |                                           |     |
|     |                                           |     |
|     | 59                                        |     |
|   2 +                                           +     |
|     | 27    50    74                            |     |
|     | 28    54    61    63    87                |     |
|     | 46    48    66    98                      |     |
|   1 + 76    85                                  +     |
|     | 4     40    41    5     65    86    91    |     |
|     | 10    11    12    13    14    35    6     |     |
:     : 60    7     84    88    90                :     :
|     | 16    17    2     26    30    32    34    |     |
:     : 55    57    67    8     9     92          :     :
*   0 * 19    37    51    70    89    99          * --- *
:     :                                           :     :
|     | 1     18    3     36    39    83          |     |
|     | 103   104   106   107   108   15    33    |     |
:     : 38    47    56                            :     :
|     | 102   21    25    29    31    53    78    |     |
:     : 82    93    96                            :     :
|  -1 + 101   105   58    79    80    81    94    +     |
:     : 95    97                                  :     :
|     | 100   23    43                            |     |
|     | 20    44    45    69                      |     |
|     | 22                                        |     |
|  -2 + 42    49                                  +     |
|     |                                           |     |
|     |                                           |     |
|     | 75                                        |     |
|  -3 +                                           +     |
|-----+-------------------------------------------+-----|
|Measr|+Ratee                                     |Scale|
+-------------------------------------------------------+
```

Figure 2. Wright Map of Ratees' Competency Level (N=100)

The total number of ratees is 100, but only 93 ratees have stated complete demographic information on gender. Therefore, the competency level and gender comparison are only based on the 93 ratees. The high-ability and low-ability ratees categories consist of both genders. These findings showed that the ratee's gender does not affect the ratee's competency level.

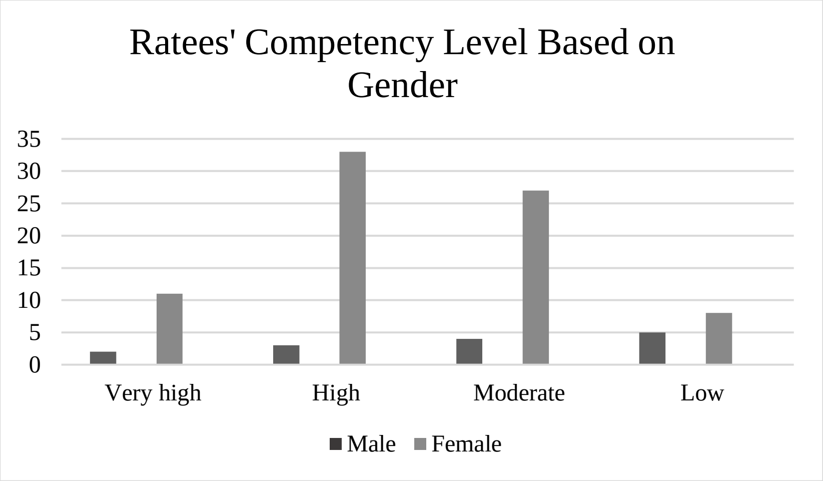


Figure 3. Comparison between competency level and gender

Statistical Inference

Independent-samples t-test was conducted to test the significant differences *in competency level between male ratees and female ratees.* In addition, the researchers did homogeneity testing using Levene's test before the difference test. First, the researchers looked at the significance level of Levene's test. If Levene's test's significant value (p) is more than 0.05, the researchers should use the first line in the table (equal variances assumed). If the p-value is equal to or less than 0.05, the variances for the two groups (males/ females) are not the same.

As a result, the data contradict the equal variance assumption. SPSS offers an alternate t-value to deal with the fact that the variances are not equal. The information in the second line of the t-test table, which refers to equal variances not assumed, should be used by the researchers (Pallant 2011).

Table 6. The Findings of the Independent-Samples Test

Independent Sample Test						
		Levene's Test for Equality of Variances				
		F	Sig	t	df	Sig (2-tailed)
Logit	Equal variances assumed	8.453	0.005	-0.760	98	0.449
	Equal variances not assumed			-0.511	16.710	0.616

Based on the SPSS output, the value of Levene/ Test for Equality of Variances was p = 0.05. So, it shows that the data variance between male ratees and female ratees was not homogeneous or the same, so the interpretation of the output independent samples t-test was guided by the values contained in the equal variances not assumed column. Based on the SPSS output, it could be seen that the significant (2-tailed) at the equal variances assumed was *-0.511*. There was no significant difference in scores for males (M = -0.18, SD = 1.64) and females (M = 0.035, SD = 0.88; t = -0.511, p = 16.71, two-tailed). Therefore, the null hypothesis (Ho1) is accepted. There is no significant difference in competency level between male ratees and female ratees. These results indicated that the level of competency for male and female ratees are the same.

Fit Statistics of Ratees

The statistical analysis of ratee aims to determine the compatibility of the data with the Rasch model based on the MnSq outfit value. To ensure that the data fit the Rasch measurement model, the researchers check the value of the outfit MnSq for each ratee. MnSq = 1 shows ideal data according to the Rasch specification. The acceptable value for its statistic ranges from 0.5 to 1.5 (Bond & Fox, 2015).

MnSq is a square mean fit statistic that determines the randomness of a measurement system. As a result, this study conducted a fit statistical analysis to guarantee that the ratees were compatible with the MFRM. The optimum MnSq value is 1.00 logits, which represents the expected value; a value less than 0.5 indicates that the data collected is easy to predict (data overfit model), and a value larger than 1.5 indicates that the data collected is difficult to predict (data underfit model) (Azrilah et al., 2013; Bond & Fox, 2015).

Table 7. Fit Statistics Analysis Findings (Misfit Ratees)

Ratee	Model		Infit		Outfit		Correlation
	Measure	S.E.	MnSq	Zstd	MnSq	Zstd	PtMea
81	-0.95	0.26	0.19	-5.10	***0.12***	-5.92	0.96
103	-0.56	0.31	0.18	-4.41	***0.12***	-5.01	0.97
100	-1.13	0.30	0.26	-3.72	***0.22***	-4.01	0.17
37	-0.10	0.25	0.35	-4.02	***0.24***	-4.01	0.92
101	-0.97	0.37	0.27	-2.93	***0.25***	-3.04	0.13
82	-0.79	0.21	0.56	-3.80	***0.26***	-4.78	0.60
94	-0.97	0.26	0.49	-2.92	***0.28***	-3.87	0.93
34	0.32	0.30	0.29	-3.59	***0.30***	-3.31	0.05
102	-0.75	0.21	0.58	-3.62	***0.31***	-4.22	0.91
67	0.22	0.37	0.39	-2.26	***0.34***	-2.42	0.15
106	-0.42	0.20	0.61	-3.30	***0.37***	-3.81	0.59
12	0.60	0.27	0.58	-2.06	***0.38***	-2.97	0.30
78	-0.68	0.25	0.60	-2.07	***0.43***	-2.96	0.36
30	0.30	0.26	0.47	-2.66	***0.43***	-2.82	0.13
99	0.00	0.36	0.40	-2.24	***0.45***	-1.44	0.98
40	0.85	0.33	0.80	-0.72	***0.48***	-1.97	0.86
1	-0.22	0.17	0.77	-2.94	***0.49***	-3.35	0.76
56	-0.60	0.19	1.56	3.78	***1.53***	2.62	0.70
20	-1.43	0.15	1.38	5.22	***1.56***	5.11	0.55
58	-1.12	0.18	1.30	3.30	***1.59***	1.80	0.82
49	-2.05	0.16	1.61	5.71	***1.68***	4.52	0.30
83	-0.23	0.16	1.54	5.42	***1.72***	5.37	0.59
41	0.81	0.20	1.24	1.96	***1.96***	3.79	0.36

The findings showed 17 ratees with MnSq outfit values less than 0.5, and there were six ratees with MnSq outfit values greater than 1.5. In addition, the findings also showed that the standard error is small, which is less than 0.50, which indicates the accuracy of the measurement. A standard error value of less than 0.25 is considered excellent (Fisher, 2007). To ensure ratees were parallel in construct measurement, the researchers examined the PtMea correlation values. The findings showed that the PtMea correlation values obtained by all ratees are positive. It proved that all ratees parallel with the measured construct. In total, 23 out of 100 ratees had outfit values that did not meet the acceptable range.

Unexpected Response

One of the strengths of MFRM is that it can provide information about the function of the elements involved in an unexpected response, such as if the raters are having trouble understanding and implementing the criteria (Eckes, 2015; Kudiya et al., 2018). The findings indicated that 64 responses showed that the rater gave a lower rating than expected (under-value), while 34 showed the rater gave a higher rating than expected (over-value).

The number of unexpected responses discovered was far too low, at only 0.48% (100 out of 21000), showing that all raters had made a comprehensive and careful judgement. In addition, unexpected responses accounted for 42.00% (42 out of 100 ratees). The frequency of each ratee discovered in unexpected responses is shown in Table 6. Only 11 misfit ratees (20, 30, 37, 40, 41, 58, 67, 78, 83, 99, and 106) out of 42 ratee appeared in an unexpected response. Ratee 41, which produced the most unexpected responses, was a misfit ratee. This study suggests that the findings of unexpected responses are not clear evidence of the misfit item. In addition, this study also shows that MFRM analysis is sensitive to detect if the ratees are easy to assess by raters or confuse the raters. When making a decision, these 42 ratees baffled the rater since they appeared in unexpected responses compared to other ratees.

Table 8. Summary of Unexpected Response Analysis Findings

Ratee	Frequency
4, 7, 17, 18, 19,22, 28, 31, 37, 40, 44, 47, 67, 75, 76, 78, 90, 98, 104, 106, 107	1
15, 20, 21, 25, 36, 58, 108	2
8, 13, 30, 32, 35, 38, 99	3
16, 85, 87	5
33, 83, 84	6
41	11

DISCUSSION

This article aims to analyse the mathematics teachers' competency level using the MFRM approach. The data analysis in the study shows that it fits the Rasch model (Table 2), with a principal component analysis of residuals of more than 40%, indicating that the instrument utilised has good unidimensionality (Andrich & Marais, 2019; Liu & Lim, 2020). This study suggests that using a multi-rater strategy to examine latent variables of ratees' classroom assessment, three constructs with 56 items of the instrument function very well (Bond & Fox, 2015; Zuliana et al., 2021).

Furthermore, all reliability indices (reliability, strata, and separation) exhibit outstanding results, indicating a multi-rater approach situation in which volume data increases compared to self-administered data (Eckes, 2015; Engelhard & Wind, 2018). Overall, the findings demonstrated that, compared to another measurement model, the MFRM could thoroughly analyse the instrument's reliability and validity in multi-rater contexts and detail (Boone et al., 2014; Eckes, 2015; Engelhard & Wind, 2018).

Another helpful feature of the MFRM is detecting ratee inconsistency in unexpected response analysis. The data screening process removes respondents who are outliers and do not match the model is very important to ensure that statistical analysis findings are valid (Widhiarso & Sumintono, 2016). Further, the study's findings also analyzed fit statistics for the ratees, which show their quality work. The findings detected that 42 ratees appeared in unexpected responses analysis, 42.00% of the total showing most ratees were easy to assess by raters. They didn't confuse the raters when making the judgement. The findings on unexpected responses demonstrated the benefits of MFRM in providing evidence for multi-rater quality assessment and ensuring more accurate and precise measurement measurement (Andrich & Marais, 2019; Bond & Fox, 2015). This study demonstrates that MFRM can identify unexpected responses, implying that improved analysis can be achieved (Engelhard & Wind, 2018).

Besides that, this study used the mean and standard deviation of the ratee's logit to categorise the ratees' competency levels into four groups (Figure 2). In this study, 16% of ratees were very high ability, 37% of ratees were high ability, 33% of ratees were moderate ability, and 14% of ratees were low ability. *Furthermore, this study revealed no significant difference in competency level between male ratees and female ratees. These results indicated that the level of competency for male and female ratees are the same. The study by Uvie (2021) found no significant difference in secondary school teachers' competency between male and female teachers, although there are differences in teaching qualifications and experience among the teachers.*

The study by Nurul Syahada (2017) found that the teacher's competency level based on gender shows no difference in perception between males and females. Gender and cumulative grade point average had no significant impact on teacher candidates' attitudes (Ozan & Kincal, 2017). Teacher assessment knowledge does not differ significantly based on gender factors, positions, teaching experience, and subject fields (Rohaya & Mohd Najib, 2008). *Gender was an ineffective independent variable in determining teacher competency in classroom assessment. Further research should be focused on other independent variables that affect a teacher's competency. Curriculum designers use assessment results to evaluate the curriculum's effectiveness based on the quality of student learning experiences, content, and the recommended assessment approach (Hussain et al., 2021) . It is found that teachers still have weaknesses in implementing CA* (Sh. Siti Hauzimah, 2019). *However, some studies showed that teachers have an excellent readiness to implement CA* (Yuh & Kenayathulla, 2020) *and have high levels of CA implementation* (Sh. Siti Hauzimah, 2019). *The study by* Al-Bahlani (2019) *showed teachers' competency level was moderate and proposed organizing a readiness program to guide teachers to enhance the effectiveness of the assessment.*

Before the introduction of the CA, teachers were found to be lacking confidence in conducting school assessments due to a lack of training, knowledge, and skills despite taking assessment courses (Fakhri & Mohd Isha, 2016). The teacher's competency in assessment is valuable because assessment is inseparable from teaching and learning. Teachers who have received courses or training on assessment tend to obtain a higher level of competency in the CA than teachers who do not attend courses or training on assessment (Murukutla, 2019). This statement is similar to the study by Sartaj et al. (2019), which found that if the teacher is given appropriate training in the assessment technique, it will benefit teachers and students.

Although CA was first introduced in 2018, the individual's ability level may have been influenced by the courses or training related to CA. The CA training benefits the teachers involved and enhances the willingness of the teacher to implement CA even though the CA is a new challenge for teachers (Kannan et al., 2021). Teachers who have attended training programs during their pre-service are expected to create effective assessments to enhance student learning (Anam & Putri, 2021). In service, teachers should also be given courses or training that focus on the CA method, procedure, purpose, planning, and implementation (Khanna & Talwelkar, 2021). Lack of effectiveness in courses or training can harm the implementation of the CA by the teacher (Zahari et al., 2020).

The effectiveness of the CA contributes to improving student performance, evaluating teacher success, strategies, and teaching methods, which in turn contributes to the overall improvement of the education system's progress. Besides the professional considerations, implementing the CA should follow the guidelines (Halimah & Rozita, 2019; Sh. Siti Hauzimah, 2019). Teachers need to increase their mastery and understanding of information as they can influence their competencies to implement the CA. The information delivery by stakeholders should be made continuously, integrated, widely, and clearly to overcome the constraints involved. This effort can also prevent the implementation of the CA from furthering the early action. Therefore, the implementer of the teaching program is responsible for the initial preparation to produce competent teachers and the parties involved in the professional development (Campbell, 2013).

In addition, teachers were still confused and lack of readiness to implement CA (Sh. Siti Hauzimah, 2019). Teachers who are less competent in the CA may be due to insufficient training, not using proper assessment methods or being unable to interpret data correctly (Murukutla, 2019). As the CA is one of the new things introduced in Malaysia, there is plenty of room to enhance teacher competency in the CA further. The parties need to organize more training programs for in-service teachers to strengthen the effectiveness of teachers' skills in assessment (Nyanjom, Yambo, & Ongunya, Raphael, 2020).

It is found that continuous monitoring and guidance can help teachers improve their understanding and implement the CA appropriately (Arumugham, 2020). The teacher's satisfaction level in implementing the CA is not high (Chee & Sern, 2019; Sartaj et al., 2019). Stakeholders need to pay attention to organizing courses or training related to CA to increase teachers' competency in CA, thus influencing the effectiveness of the CA. Courses or training related to the CA can guarantee the effectiveness and validity of the CA implemented (Sh. Siti Hauzimah, 2019).

Competency Theory also states that skills and knowledge are easily acquired through courses and training and can be influenced by academic qualifications (Spencer & Spencer, 1993). One advantage of competencies is that they can be developed through various training and education programs (Boyatzis, 2008). The statement was also supported by Winterton dan Winterton (1999), who stated that individual competencies could be developed consistently by the organization by enhancing knowledge, understanding, and skills. Besides, the aspects of attitude are unique features that are difficult to achieve (Spencer & Spencer, 1993). Still, attitudes can be formed, built, and changed despite complicated processes and determination (Che Ghani et al., 2018). Overall, the findings of this study contributed to the development of teacher professionalism.

CONCLUSION

The results of the multi-rater methods study using MFRM provide interesting outcomes and detailed information on the ratees' competency levels. This study also demonstrates that assessing teacher ability is difficult, but that MFRM is a great method. MFRM produces more precise information about the pattern of the ratees' ability and enhances the validity process compared to the CTT technique, which focuses on group-centered statistics (Mohd Zabidi et al., 2021). Overall, the study found MFRM to be a more effective psychometric process for assessing ratees' ability than CTT's method because MFRM is broader and gives a complete analysis of the ratees' ability (Eckes, 2019).

Furthermore, Rasch measurement model analysis produces better and more precise measurements that help the consistency in response to the questionnaire (Adams et al., 2020). Besides, this study shows how the researchers can ensure that accurate and fair measurement are produced if the multi-rater method is used. The psychometric testing for the instrument in this study was conducted using MFRM by the multi-rater method approach used. This study has also shown the advantages of MFRM compared to the CTT method to analyse research data using multi-rater methods. Multi-rater procedures can produce more fair and accurate performance measurements.

REFERENCES

Adams, D., Tan, M. H. J., & Sumintono, B. (2020). Students' readiness for blended learning in a leading Malaysian private higher education institution. Interactive Technology and Smart Education.

Al-Bahlani, S. M. (2019). Assessment literacy : A study of EFL teachers ' assessment knowledge, perspectives , and classroom behaviors. The University of Arizona.

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

Anam, S., & Putri, N. V. W. (2021). How literate am i about assessment: Evidence from Indonesion EFl pre-service and in-service teachers. Journal of English Education, 9(2), 151–162.

Andrich, D., & Marais, I. (2019). A course in Rasch measurement theory: Measuring in the educational, social and health sciences. Singapore: Springer Text in Education.

Arumugham, K. S. (2020). Kurikulum, pengajaran dan pentaksiran dari perspektif pelaksanaan pentaksiran bilik darjah. Asian People Journal (APJ), 3(1), 152–161. https://doi.org/10.37231/apj.2020.3.1.175

Azrilah Abdul Aziz, Mohd Saidfudin Masodi, & Azami Zaharim. (2013). Asas model pengukuran Rasch: Pembentukan skala dan struktur pengukuran. Bangi: Universiti Kebangsaan Malaysia.

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

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.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch analysis in the human sciences. New York: Springer.

Boyatzis, R. E. (2008). Competencies in the 21st century. Journal of Management Development, 27(1), 5–12. https://doi.org/10.1108/02621710810840730

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

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.

Campbell, C. (2013). Research on teacher competency in classroom assessment. In Research on classroom assessment. United States of America: SAGE Publications, Inc.

Che Ghani Che Kob, Mohd Zaini Osman, & Nur Faeeza Abd Ghafar. (2018). Competence of instructor in practicing teaching of furniture manufacturing in Malaysia. UTM Press Sains Humanika, 10(3), 25–32.

Chee, C. S., & Sern, L. C. (2019). Tahap kepuasan guru terhadap Pentaksiran Berasaskan Sekolah (PBS): Perbezaan persepsi dalam kalangan guru bagi mata pelajaran Kemahiran Hidup Bersepadu. Online Journal for TVET Practitioners, 4(2), 30–34.

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

DeLuca, C., Valiquette, A., Coombs, A., LaPointe-McEwan, D., & Luhanga, U. (2018). Teachers' approaches to classroom assessment: A large-scale survey. Assessment in Education: Principles, Policy and Practice, 25(4), 355–375. https://doi.org/10.1080/0969594X.2016.1244514

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

Fakhri Abdul Khalil, & Mohd Isha Awang. (2016). Isu kesediaan guru dalam amalan melaksanakan pentaksiran berasaskan sekolah. EDUCATUM – Journal of Social Science, 2, 1–7.

Fisher, W. P. . (2007). Rating scale instrument quality criteria. Rasch Measurement Transactions, 21(1), 1095.

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: Implications for measurement courses. Measurement and Evaluation in Counseling and Development, 36(3), 181–191. https://doi.org/10.1080/07481756.2003.11909741

Gunal, Y., Usta, G., & Uluman, M. (2015). An investigation of attitudes of candidate teachers towards measurement and evaluation lesson against certain variables. Procedia - Social and Behavioral Sciences, 177, 209–212. https://doi.org/10.1016/j.sbspro.2015.02.388

Halimah Jamil, & Rozita Radhiah Said. (2019). Pelaksanaan penskoran pentaksiran lisan bahasa melayu dalam pentaksiran bilik darjah. Jurnal Pendidikan Bahasa Melayu - JPBM, 9(2), 25–36.

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

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

Hussain, S., Idris, M., & Akhtar, Z. (2021). Perceptions of teacher educators and prospective teachers on the assessment literacy and practices. Gomal University Journal of Research, 37(1), 71–83.

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.

Kannan, B., Pillai, R. V., & Kunhikannan, S. K. (2021). Keberkesanan pelaksanaan bengkel pentaksiran bilik darjah. Jurnal Penyelidikan Dedikasi, 19(1), 51–72.

Khanna, A., & Talwelkar, A. S. (2021). An analysis of classroom assessment practices of english language teachers. Research Journal of English and Literature, 9(3), 288–294. https://doi.org/10.33329/rjelal.9.3.288

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

Kursad, M. S. (2022). Gender effect on competency perceptions on measurement and evaluation : A meta-analysis study. Teacher Education and Instruction, 3(1), 38–50.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. Journal of Applied Measurement, 3(1), 85–106.

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

Liu, V. Y. Y., & Lim, S. M. (2020). A psychometric evaluation of the brief resilience scale among tertiary students in Singapore. Asia Pacific Journal of Education, 1–14. https://doi.org/10.1080/02188791.2020.1845120

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

Maryati. (2019). Multirater assessment to teacher professionalism based on pedagogical content knowledge. Journal of Physics: Conference Series, 1233. https://doi.org/10.1088/1742-6596/1233/1/012085

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

Muhd Khaizer Omar, Farah Nadia Zahar, & Abdullah Mat Rashid. (2020). Knowledge, skills, and attitudes as predictors in determining teachers' competency in Malaysian TVET institutions. Universal Journal of Educational Research, 8(3C), 95–104. https://doi.org/10.13189/ujer.2020.081612

Murukutla, M. (2019). The effects of background, classroom assessment competence, self-efficacy, and self-perceived assessment skills on classroom assessment practices of teachers in India. University of Nevada, Las Vegas.

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 Mohd Radzi. (2017). Pembinaan dan pengesahan instrumen pentaksiran prestasi standard awal pembelajaran dan perkembangan awal kanak-kanak. Universiti Malaya.

Norzetty Md Zahir, & Sumintono, B. (2017). Perceptions on influence tactics among leaders in the ministry of education Malaysia : An application of the many facets Rasch model. International Conference On Public Policy, Social Computing And Development (ICOPOSDEV), (October), 1–13.

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

Nurul Nadia Abd Latib, Shahrir Jamaluddin, & Sumintono, B. (2018). Analisis multi-rater pelajaran pendidikan islam pada ujian Pentaksiran Tingkatan Tiga (PT3) di Malaysia 1 analisis multi-rater pelajaran pendidikan islam pada ujian pentaksiran. 1st National Conference on Educational Assessment and Policy (NCEAP).

Nurul Syahada Mohd Suhaimi. (2017). Standard guru Malaysia dalam program persediaan guru reka bentuk dan teknologi. Universiti Tun Hussein Onn Malaysia.

Nyanjom, A. O., Yambo, J. M. O., & Ongunya, Raphael, O. (2020). Influence of teachers' assessment competency on pupils' academic achievement in Kisumu County. Journal of Advances in Education and Philosophy, 4(11), 483–493.

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

Ozan, C., & Kincal, R. Y. (2017). An investigation of teacher candidates ' attitudes towards educational measurement in terms of various variables an investigation of teacher candidates attitudes towards educational measurement in terms of various variables. Turkish Journal of Teacher Education, 6(1), 18–32.

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

Rohaya Talib, & Mohd Najib Abd Ghafar. (2008). Pembinaan dan pengesahan instrumen bagi mengukur tahap literasi pentaksiran guru sekolah menengah di Malaysia. Seminar Penyelidikan Pendidikan Pasca Ijazah 2008, 25-27 November 2008, Universiti Teknologi Malaysia. https://doi.org/10.4103/nah.NAH_106_16

Sartaj, S., Kadri, S., Shah, S. F. H., & Siddiqui, A. (2019). Investigating the effectiveness of classroom based assessment on ESL teaching strategies and techniques in Pakistan: Study from teachers' perspective. Theory and Practice in Language Studies, 9(7), 826–834. https://doi.org/10.17507/tpls.0907.12

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

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

Sh. Siti Hauzimah Wan Omar. (2019). Pengetahuan, kemahiran, sikap dan masalah guru dalam melaksanakan pentaksiran bilik darjah bahasa melayu di sekolah rendah. 9(1), 56–67.

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.

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

Widhiarso, W., & Sumintono, B. (2016). Examining response aberrance as a cause of outliers in statistical analysis. Personality and Individual Differences, 98, 11–15.

Winterton, J., & Winterton, R. (1999). Developing managerial competence. London: Routledge.

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

Wu, S. M., & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: The case of a university placement test. Higher Education Research and Development, 35(2), 380–394.

Yuh, T. J., & Kenayathulla, H. B. (2020). Pentaksiran bilik darjah dan prestasi murid sekolah jenis kebangsaan cina di Hulu Langat, Selangor. Jurnal Kepimpinan Pendidikan, 7(3), 53–64.

Zahari Suppian, Nor Hasnida Che Md Ghazali, Nor Junainah Mohd Isa, & Govindasamy, P. (2020). Penilaian kendiri guru pelatih terhadap tahap kemahiran pentaksiran bilik darjah (PBD). Jurnal Dunia Pendidikan, 2(4), 98–106.

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.

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