Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

RATER SEVERITY IN MEASURING TEACHERS' COMPETENCY IN CLASSROOM ASSESSMENT

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

[1]Faculty of Education, University of Malaya (UM), Malaysia,
(E-mail: rosyafinazmohamat@gmail.com)
[2]Faculty of Education, University of Malaya (UM), Malaysia,
(E-mail: harris75@um.edu.my)
[3]Faculty of Education, Universitas Islam Internasional Indonesia, Indonesia,
(E-mail: bambang.sumintono@uiii.ac.id)

Article history
Received date : 15-6-2022
Revised date : 16-6-2022
Accepted date : 30-6-2022
Published date : 30-6-2022

To cite this document:
Rosyafinaz, Harris Shah & Sumintono (2022). Rater Severity In Measuring Teachers' Competency In Classroom Assessment. *Journal of Islamic, Social, Economics and Development (JISED)*, 7(45), 141 - 155.

Abstract: *Many Facet Rasch Model (MFRM) provided more honest information on the validity of the judgement process by considering the factors of raters' variability and estimating the ratee abilities without depending on the severity or leniency of raters. The multi-rater method produced a more stable and precise assessment and has higher reliability than the self-assessment method. However, the raters have different characteristics that can influence their severity level when evaluating the ratees. This study aims to investigate the effect of gender and job position on rater severity when assessing teachers' competency in Classroom Assessment (CA) by using MFRM. The instrument consists of 56 items built based on three main constructs; knowledge in CA, skills in CA, and attitude towards CA. This study used a quantitative multi-rater approach using a questionnaire distributed to 262 raters to assess 100 ratees. Raters' severity levels form four categories based on the logit measure value. There are no significant mean differences between the rating of male and female raters. Similarly, there is no significant mean difference between the rating based on job positions held by the raters.*

Keywords: *Rater Severity; Many Facet Rasch Model; Classroom Assessment; Competency; Multi-rater Analysis*

Introduction

Classroom Assessment (CA) involves gathering information by teachers to decide on the follow-up actions that need to be taken to improve the development of student learning (Bahagian Pembangunan Kurikulum, 2019). Teachers are the implementing agents responsible for ensuring the effectiveness of the assessments implemented. Teachers who perform an assessment and provide feedback correctly can be a factor in improving student performance (Lam, 2019; Sartaj et al., 2019).

Teachers will be more effective in carrying out their tasks if they have a competency that includes knowledge, skills, and attitudes (Muhd Khaizer et al., 2020; Sh. Siti Hauzimah, 2019). Competency is defined as the ability owned by an individual (Boyatzis, 2008). Competency

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

refers to proficiency, accuracy, expertise, and skill (Eraut, 2003). The low level of teacher competency is one of the causes of low student competency because teachers cannot apply the quality of teaching and learning (Permana, 2017).

CA issues also require teachers to be competent to assess and make judgments about student achievement (Hariatul et al., 2021). The study of teachers' competency in assessment is critical because it is an effort toward effective and relevant assessment to preserve the quality of the education system (Uvie, 2021). If the study is not implemented, it may disrupt the government's intention to implement the holistic assessment to improve the teacher's quality. Using the rater's judgement to determine the ratee's competency can affect the ratee's final performance (Engelhard & Wind, 2018).

The teachers' competency issues have attracted many researchers and raised concern among many stakeholders in Malaysia due to the emergence of question marks about how far the teachers educate their students (Muhd Khaizer et al., 2020). It is essential to determine the teachers' competency level because it involves long-term effects, perception, and making judgments in executing tasks (Foschi, 2000). The instruments for measuring teachers' competency guide their education programs and help them reflect on their competency level towards the assessment (Stiggins, 1999). The lack of research on competency in the context of in-service teachers in Malaysia has raised questions about the instrument to measure teachers' competency in implementing CA in the local context.

Furthermore, most previous studies used self-assessment to determine teachers' competency, depending on respondents' honesty. The self-assessment cannot accurately measure the respondent's behaviour, even though the information obtained is helped provide diagnostic information (McMillan, 2013). Self-assessment may be limited by the individual's willingness to share truthful information, the presence of an element of bias, and their ability to make an appropriate self-assessment (Ganellen, 2007). Self-assessment is subjective, causing an individual to give a lower or higher response than their actual ability and does not describe their proper behaviour (bias). The reason is that self-assessment depends on respondents' honesty and does not necessarily reflect the actual behaviour.

This study involves a multi-rater method that does not rely on self-assessment. In this study, the raters used the instrument to assess the teachers. The researchers should not underestimate the problems regarding the raters' assessment. Raters who fail to control their severity level can influence the difference between the observed and expected scores, thus negatively affecting ratees performance (Muhammad Firdaus & Mohd Effendi, 2020). This result may negatively affect the ratee's competency measurement (Bond & Fox, 2015). The multi-rater method causes some teachers to be judged by severe raters and some teachers to be evaluated by lenient raters. There is also the possibility that some raters are inconsistent when judging. This study investigates the effect of gender and job position on rater severity when assessing teachers' competency in Classroom Assessment (CA) using MFRM.

Literature Review

Teachers' Competency Influenced by Variability among Raters

There is a lack of empirical studies on teachers' competency in CA (Murukutla, 2019), indicating the need for quality instruments to measure the teachers' competency level in CA. Most of the past studies about the teacher assessment have focused more on pre-service

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

teachers, but less attention on in-service teachers (Campbell, 2013; DeLuca et al., 2016). Although there are instruments built related to teachers' competency in the country, most of these instruments were only focusing on the basic aspects of teachers' competency and lack of focus on the competency theory which also includes the aspects of teacher personal quality (Zahari, 2018).

Furthermore, the reliability of the teacher evaluation is higher when it involves multiple raters (Kane & Staiger, 2012). The peer assessments can increase the reliability and validity of assessments and suitable for the evaluation of aspects in workplace (Schmidt et al., 2016). Multi-rater methods have become increasingly popular, involving peer assessment, self-assessment, and the assessment of superiors or subordinates to determine an individual's job performance (Scullen et al., 2000). Evaluating teachers' quality and performance is recommended to involve more than one rater because it is often seen as the 'key' to successful teacher evaluation practices (OECD, 2013). Raters' different characteristics can influence their severity level when evaluating the ratee (Myford & Wolfe, 2003).

Discussion on rater's judgement quality is very important to ensure that ratees are judged with fairness and reliability (Muhammad Firdaus & Mohd Effendi, 2020). The consistency of the raters in performing the assessment has attracted many previous researchers, especially in the education, language and psychology area (Engelhard & Wind, 2018). Rater's assessment quality can be affected by the rater's variability that can prevent them to produce a valid and reliable score, which may not be an accurate measurement to determine the ratee's competency (Wu & Tan, 2016).

In performance assessment, raters tend to have different tendencies, errors, and biases known as rater effects, threatening the judgment's validity and fairness (Jin & Eckes, 2021; Myford & Wolfe, 2003; Saal et al., 1980). The raters' effect occurs not because of the measurement object but is caused by the rater and can affect the study (Engelhard & Wind, 2018; Styck et al., 2020). The raters' effect may indicate an inappropriate or unfair assessment that has been made (Styck et al., 2020). The process of data screening to eliminate outliers and misfit respondents is very reasonable to ensure that the statistical analysis results are valid (Widhiarso & Sumintono, 2016).

In the context of MFRM, a rater severity is defined as the tendency of raters to consistently give a lower score than the other raters (Myford & Wolfe, 2004). There is a possibility that the raters' severity level is influenced by various factors, such as the difference in 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 raters' judgement (Eckes, 2015). It is essential to examine the raters' behaviour and correct the sources of error to ensure the validity and reliability of the measurement (Aslanoglu & Sata, 2021).

The previous studies found that demographic factors such as age, gender, and experience can influence the raters' judgement (White et al., 2002). The studies on writing assessments identified a significant effect of the interaction between raters' gender (Gyagenda & Engelhard, 2009). Female raters tend to be more consistent than male raters when judging (Barth & Stadtmann, 2019). Expert and non-expert raters may have different severity levels when making judgments (Barth & Stadtmann, 2020).

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

But, although the raters have different backgrounds in terms of age, gender, and education, there may be no significant differences between personality traits and the severity level of the rater (Esfandiari, 2019). There was no significant relationship between teaching experience and teachers' competency level in CA (Murukutla, 2019). In addition, there were no significant differences between the teachers' implementation level in CA and the years of teaching experience (Norshafinaz & Faridah, 2018; Yuh & Kenayathulla, 2020). The study by Chee and Sern (2019) also showed no significant differences in the satisfaction level of the assessment implementation based on the teachers' experience.

A detailed scoring procedure is essential to produce an appropriate score to measure the construct. Judgment by the outliers may cause unreliable measurements, distort the relationship between the variables studied, and cause problems in interpretating the analysis (Widhiarso & Sumintono, 2016). Therefore, various psychometric models and statistical indices are used to ensure the high quality of the assessment (Engelhard & Wind, 2018; Wind & Peterson, 2018). The aberrant responses are caused by several factors, such as the inconsistent response, a response that consists of extreme score values, or a response that only gives the same score for all items (Widhiarso & Sumintono, 2016). The aberrant response patterns can be avoided using Rasch model analysis to ensure that the responses fit the measurement model (Panayides & Tymms, 2013). The Rasch model analysis yields better and more accurate measurements to obtain the consistency of questionnaire responses (Adams et al., 2020).

The Problem of Detecting Rater Severity in Multi-rater Setting in Classical Test Theory

Three main factors influence the assessment of an individual's achievement: the ratee's actual achievement, rater bias, and random measurement error (Wherry & Bartlett, 1982). This study uses the multi-rater method to overcome the existence of bias assessment, which involves self-assessment and several raters with the experience and expertise to judge the teachers' competency level. The multi-rater method is appropriate to ensure that the judgement is honest and fair.

Various statistical methods in the Classical Test Theory (CTT) approach have been widely used to analyze the data involving the raters' judgement. However, the CTT method is not ideal because it does not provide detailed information on item difficulty, rater severity, and difficulty of the dimensions for each facet (Crocker & Algina, 2008). The statistical method approach used in the data analysis of the multi-rater method can also raise questions.

Moreover, the CTT method cannot identify differences between raters, such as determining if the raters have a consistent severity level in the judgement (Newton, 2009). Generalizability Theory (G theory) is a method developed to overcome the limitations of the commonly used CTT method. Nevertheless, it is found that the G theory makes it difficult for the reader to understand the interpretation because the method is quite complicated and complex (Brennan, 2010; Webb et al., 2018). Another limitation found in the G theory and Fuzzy Delphi methods is the inability to identify the raters' severity level, contributing to the explanation of the rating scale without considering the raters' severity error (Zhu et al., 1998).

A study by (Scullen et al., 2000) used a multi-rater method, showing the CTT analysis was unable to provide details about the psychometric characteristics of the item, the ability level of the ratee, and the consistency of the raters' judgement. Moreover, the study's findings provided only 21% of information about the ratee and 62% about the raters' judgment. Therefore, this

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

study shows how researchers can ensure that the multi-rater method can produce accurate and fair measurements using MFRM analysis.

MFRM in Research

MFRM is a sequel of the Rasch model and involves more than two interacted aspects to generate observation (Linacre, 1994). The MFRM can combine more variables or facets to determine the relationship between these facets, for example, an analysis involving three facets, namely item, rater, and ratee (Eckes, 2015). MFRM analysis using the Facets software can provide detailed information because it can report the statistical results for each facet (Sudweeks et al., 2005). Another advantage of MFRM is that each raters' judgement is based on its assessment style, which is not influenced by other raters (Bond & Fox, 2015; Engelhard & Wind, 2018). The MFRM has been widely used in various fields because it is more practical to meet the needs of the validity and use of the assessment results to improve the development of methodologies in future studies (Eckes, 2015).

The Rasch model helps researchers review instruments (add or remove items), detect possible biases in measurements, and make it easier for researchers to communicate the findings, such as using Wright Maps to make precise comparisons of individual ability and item difficulty (Boone, 2020). MFRM has the advantage to model the raters based on their own definition of the scale, without having a parallel judgement with the other raters (Bond & Fox, 2015; Eckes, 2015; Engelhard & Wind, 2018). The raters' bias that exists based on the raters' severity level is defined as the raters' behaviour that often applies in the performance assessment process and may affect the validity of the assessment (Erman Aslanoglu, Karakaya, & Sata, 2020).

Many studies used MFRM to explain the raters' effect in judging the ratees' performance, such as the raters' tendency to be severe or lenient (Eckes, 2005; Farrokhi et al., 2012; Lumley & Mcnamara, 1995; Schaefer, 2008). A study by Maryati (2019) using the multi-rater method to judge a teacher's professionalism towards pedagogical content knowledge found that MFRM provided clear information regarding the abilities of 20 teachers assessed by six raters, indicating that MFRM may produce appropriate analysis using a small sample. The findings are similar to a study by Nurul Nadia et al. (2018), which stated that the multi-rater method could produce a more precise assessment because it can avoid bias and the MRFM analysis attempts to place the individual responses, items, and assessment on the same interval scale.

A study by Muhammad Firdaus and Mohd Effendi (2020) used the MFRM to identify the rating performance among raters in assessing oral tests among secondary school students. The study involved 30 respondents consisting of English teachers. The findings showed that the raters have different severity levels when doing the judgement. The results also found no significant differences between the rating performance of inexperienced and experienced raters. The study also suggested that further studies can include the effects of experience and interaction effects to evaluate the rating performance among raters. In addition, this study can inspire other researchers to obtain more accurate measurements.

The use of MFRM is increasingly being used in studies that involve multi-rater methods. For example, a study conducted by Springer and Bradley (2018) used MFRM to determine the influence of raters on the assessment of a live concert band festival. The results reveal that raters play an essential role in the judgement. The study by Wang et al. (2021) was one of the first research in the Canadian context using MFRM, referred to by Canadian Language Benchmarks (CLB). This study examined the raters' performance against the Canadian English Language

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

Benchmark Assessment for Nurses (CELBAN) components based on MFRM analysis. The study identified the raters' reliability is based on the aspects of the raters' consistency and severity, bias judgement, and the use of a rating scale. The study results showed that the raters have a consistent judgement pattern based on the raters' severity.

A study conducted by Wu and Tan (2016) using MFRM to identify rater's scoring behaviour and establish how it affects student's performance. The study showed that the raters have a different severity level when doing judgement despite training. The study also showed that MFRM can explain the rater's pattern in scoring and the analysis of MFRM can produce data that enable the researcher to handle the practical issue to manage rater differences. These indicate that MFRM is an alternative model suitable to overcome the limitations in CTT statistical models. MFRM is used in this study to obtain a fair, accurate, and precise assessment based on the rater's judgment.

Research Methodology

Instrumentation

Instrument to Measure Teachers' Competency in CA consists of 56 items to measure three main constructs: knowledge in CA, skills in CA, and attitude towards CA. The teachers' competency in CA is measured by using the multi-rater method. The total number of items for each construct is 22 items for knowledge in CA, 24 for skills in CA, and ten for attitude towards CA. The instrument constructs are developed based on analysing eight competency models and 13 existing instruments, adjusted to the Classroom Assessment Implementation Guidelines (Second Edition) from Bahagian Pembangunan Kurikulum (2019). Each item was assessed based on a 5-point Likert rating scale as response options for all the items; the higher the score, the better the performance of the ratee.

The Respondents

This study's population is Mathematics teachers serving 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 are used to identify the respondents. The cluster sampling technique was used to categorize Selangor into ten districts. Then, simple random sampling was used to select four districts, six schools for each district, and the five teachers to be assessed for each school (ratee). Finally, the purposive sampling technique was used to determine the five raters for each ratee.

The first step in the Rasch model analysis is detecting and eliminating respondents who do not match the model. This study involved 324 raters who assessed 108 teachers. In total, 57 teachers were rated by five raters, four raters rated 18 teachers, three raters rated 23 teachers, and three raters rated ten teachers. Excluded respondents did not respond to the assessment scale or gave unexpected responses, e.g., intentionally creating negligence while responding (Kreijns et al., 2018). After the data screening process, the number of respondents used in the analysis of this study was 262 raters to assess 100 teachers. 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.

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

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

In this study, the raters' severity level was determined using an analysis of MFRM. The data collected were recorded in Microsoft Excel and then analyzed using the Facets version 3.71.3 software. To ensure that the data fit the Rasch measurement model, the researcher examined each rater's value of the MnSq outfit. MnSq is a mean square statistic that determines the randomness of a measurement system (Azrilah et al., 2013). The value of MnSq = 1 indicates that the data is ideal according to the Rasch specification. A statistical range of equivalence of 0.5 to 1.5 is acceptable (Bond & Fox, 2015).

The Rasch model can convert ordinal data to interval data based on the logarithmic probability method, making it a log-linear model using logit units of measurement (Linacre, 2006). Thus, interval scales with equal distances are generated along a linear line along a logit scale. The logit measurement estimates the severity level of the rater, i.e., a large logit measure value indicates that the rater has a high severity level. In contrast, a small logit measurement value indicates the rater has a low severity level. These findings will help the researcher identify the raters' severity level based on the value of the logit measurement. The raters' separation index is also used to determine the raters' severity levels distribution (Styck et al., 2020). The separation index value that exceeds three also indicates a good representative of the rater based on the individual severity level of the rater when making the judgement.

The analysis conducted using Facets software can obtain a statistical value of the percentage of agreement between the raters, indicating the extent to which the raters agree with the given judgement. The actual agreement is the percentage of agreement for the judgement between the raters. At the same time, the value of the expected agreements is the percentage that should be achieved if the data is consistent with the Rasch model. A commonly accepted value is when the observed agreement percentage value slightly exceeds the expected agreement percentage

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

value (Linacre, 1994). The reliability values are acceptable if greater than 0.80 (Bond & Fox, 2015).

The logit measure value obtained from the MFRM analysis was recorded into the Statistical Package for Social Sciences (SPSS) regarding the raters' severity level. Next, the independent sample t-test analysis was run to test the differences in the rating of teachers' competency by gender. One way ANOVA analysis was carried out to examine the differences in rating teachers' competency by job position.

Research Findings

The Reliability of Raters

The raters' reliability value is very high, which is 0.98, and the separation index of 7.46 is good as it is above 3. The significant value is p = 0.00, indicating a significant difference in the raters' severity, i.e., there was high internal consistency in the raters' judgment. These indicated that all the raters have different severity levels when judging. In this study, the reliability analysis of the raters contributed to the findings of the local independence analysis. Such results are essential to ensure the raters make judgements without being influenced by other raters.

The researchers compare the actual percentage of agreement of raters and the percentage of expected agreement of raters to ensure that the raters have made the judgement without being influenced by other raters. The findings show that the actual percentage of agreement of raters was 52.1%, whereas the percentage for expected agreements of raters was 52.2%. The almost same percentage values indicated that the judgement made by the raters is good and meets the expectations of the Rasch model.

Table 2: **MFRM Analysis Findings**

	Value
N	
Mean Logit	-3.71
Standard Deviation (SD)	3.14
Standard Error (SE)	0.39
Separation Index	7.46
Strata	10.28
Reliability Index	0.98
Significant (p)	0.00
Observed Exact Agreements (%)	52.1
Expected Agreements (%)	52.2

Fit statistics

The Rasch model has the advantage of identifying which raters have given inconsistent responses that are unpredictable by the model. The fit statistics results show that there are 16 outliers (minimum). The MnSq outfit values are sensitive to outliers that facilitate researchers to identify and correct the fit-related issues (Boone et al., 2014). These findings can be because some raters have given a minimum or maximum score to all items or given the same score to all items. The measurements will become weak if misfit raters and outliers are not eliminated (Linacre, 1994).

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

The findings show that there were 39 misfit raters. Eight ratees need to be eliminated for not having enough raters after the outliers and misfit raters are eliminated. Overall, the total number of responses eliminated in the data screening process was 95 responses out of 470 responses. Thus, the total number of responses remaining was 375 responses. The response was an assessment by 262 raters of 100 teachers. This number is sufficient according to the number of respondents in the Rasch Model, which requires a sample of 243 people to meet the ± 0.5 logit scale with 99% reliability (Boone et al., 2014; Linacre, 1994).

Rater Severity

The analysis showed that rater R285 is the most severe rater with a logit value of 7.77 (SE = 0.34), while rater R180 is the most lenient rater with a logit value of -12.36 (SE = 1.01). Overall, the findings indicated that raters with different levels of severity made assessments without being influenced by other raters, contributing to the high internal consistency in the raters' judgement. Based on the logit values obtained, the researcher decided to categorize the raters' severity levels in more detail (Table 3) based on the mean logit = -3.71 and the standard deviation = 3.14.

Table 3: **Category of Raters' Severity Level Based on Raters' Demographic**

Raters' Demographic	Very high severity level (Logit > -0.57)	High severity level (Logit -3.71 to -0.57)	Medium severity level (Logit -6.85 to -3.71)	Low severity level (Logit < -6.85)
Male	7 (2.67%)	6 (2.29%)	10 (3.82%)	5 (1.91%)
Female	25 (9.54%)	86 (32.82%)	89 (33.97%)	34 (12.98%)
SISC+	0 (0.00%)	0 (0.00%)	4 (1.53%)	2 (0.76%)
The Head of Mathematics & Science Department	1 (0.38%)	8 (3.05%)	6 (2.29%)	2 (0.76%)
The Head of Mathematics Panel	0 (0.00%)	9 (3.44%)	6 (2.29%)	4 (1.54%)
Mathematics Teacher	31 (11.83%)	75 (28.63%)	83 (31.68%)	31 (11.83%)

Does the rater's severity differ by gender? An independent samples t-test was performed to examine the difference. The t-test analysis was conducted to identify the differences in raters' severity levels based on gender factors by testing the statements of the following hypotheses:

H_{01} = There was no significant difference in the mean value of the raters' severity level based on the gender factor.

The findings showed the significant level of Levene's test is p=0.024 or less than 0.05, which means that the variance for the two groups (males/females) is not the same. Therefore, the data violate the assumption of equal variance. The researchers used the information which refers to equal variances not assumed. The severity level of male raters (M = 3.31, SD = 3.94) was lower than female raters (M = 3.76, SD = 3.04). The findings found that the t-value for the severity level comparison of male and female raters is t = 0.583 and the significant level of p = 0.564.

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

This significance level was more than 0.05 ($p > 0.05$). Therefore, the null hypothesis (Ho_1) is not rejected. So, there is no significant difference in the severity level between male and female raters.

The one-way ANOVA test analysis was conducted to identify the differences in raters' severity level based on job position factors by testing the statements of the following hypotheses:

H_{02} = There was no significant difference in the mean value of the raters' severity level based on the job position factor.

The severity level of SISC+ Officers (M = -6.65, SD = 1.07) was the lowest compared to the Head of Mathematics & Science Department (M = -4.01, SD = 2.28) and the Head of Mathematics Panel (M = -4.60, SD = 2.71). The severity level of Mathematics teachers was the highest. Hypothesis testing through parametric testing (ANOVA) was used to determine the significant mean differences between the categories of job positions held by the raters. The significant value of $p > 0.05$ (i.e., $p = 0.053$). Although the mean value was different for each position held, the findings suggested no significant mean difference between the rating tendency of job positions held by the raters. Therefore, H_{02} failed to be rejected.

Discussions

This study indicates that all the raters have different severity levels when judging. The study by Schaefer (2008) supports the effectiveness of MFRM in analyzing the test score variability caused by rater bias. MFRM provides more honest information on the validity of the judgement process by considering the factors of raters' variability and estimating ratee abilities independent of the severity or leniency of raters (Linacre, 1998). A previous study found that some raters tend to be more severe or lenient than others in the multi-rater approach (Lumley & Mcnamara, 1995; Shin, 2010; Wigglesworth, 1993).

Inconsistent judgement can affect the validity and reliability of performance assessment and cause the scores to be questioned (Aslanoglu & Sata, 2021; Schaefer, 2008). The fit statistics analysis used in the study by Erman Aslanoglu et al. (2020) also found that the MnSq outfit values of the raters were not in the accepted range and can identify if the raters were outliers. The data screening process that removes respondents who are outliers and misfits from the model is very reasonable to ensure that statistical analysis findings are valid (Widhiarso & Sumintono, 2016).

Although the raters were given the same rating scale and interpretation, each rater could not produce the same behaviour and assessment results (Wang et al., 2021). The raters' severity level differences apply if the raters interpret the scale category differently or have different standards and goals (Noor Lide, 2011). The differences in the raters' severity level toward specific criteria are caused by some raters considering some criteria to be very important or less critical (Eckes, 2012).

Factors such as rater background, rater mother tongue, training, rater cognition, rating scales, and rater experience may influence the raters' judgement (Barkaoui, 2011; Eckes, 2012). This study found that the raters' gender does affect raters' severity level. These findings are similar to a study by Erman Aslanoglu et al. (2020), who statistically found no significant differences in raters' severity based on gender factors. Their study aimed to determine the behaviour of university student raters based on self-assessment and peer assessment using MFRM analysis. The fact that there were no significant differences in raters' severity based on gender factors

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

suggests that raters had almost similar levels of behaviour (Erman Aslanoglu et al., 2020). Thus, the results of this study seem to be consistent with the previous literature in this area.

This study found no significant mean difference between the rating tendency of job positions held by the raters. The study by Barkaoui (2011) found that inexperienced raters tend to be more lenient than experienced raters when judging using scale analysis. But experienced raters tend to have a higher consistency than inexperienced raters (Sweedler-Brown, 1985). The study by Cigularov and Dillulio (2020) showed that employees who hold positions as supervisors and non-supervisory tend to make a different judgement.

The raters tend to have different behaviour in the judgement, which may affect the assessment results. This situation may contribute to serious raters' errors (Cronbach, 1990). A study by Erman Aslanoglu et al. (2020) showed that the MFRM analysis could identify the differences in the severity level between self-assessment and peer assessment. Their study also found that the raters showed a low severity level when making self-assessments but a high severity level when making peer assessments. In addition, their study found that self-assessment showed lower reliability based on the standard error of self-assessment, which was greater than the standard error of peer assessment.

The raters may be given courses or training to get more information about the severity level and the consequences of the assessment. The raters showed improvement in their consistency and severity in the assessment after receiving training (Davis, 2015). In addition, the raters need an explanation regarding the measured construct and the scale category used so that the raters can understand and differentiate the scale categories (Myford & Wolfe, 2003).

Conclusion

In the CTT approach, interrater reliability values are high only when the raters have a similar agreement in their judgement (Noor Lide, 2011). The use of MFRM in this study helped the researcher identify inconsistent raters. The results should not include the judgement made by inconsistent raters because they can affect the accuracy of the measurements. Thus, it showed that using the multi-rater method in an assessment could guarantee a more accurate, transparent, and fair measurement. This study showed that MFRM could provide accurate measurements and produce the necessary information in detail. The researchers can examine the rater's quality based on fit statistics and severity. Overall, MFRM can be widely used to improve the quality of the rater's assessment.

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*.

Aslanoglu, A. E., & Sata, M. (2021). Examining the differential rater functioning in the process of assessing writing skills of middle school 7th grade students. *Participatory Educational Research (PER)*, *8*(4), 239–252.

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.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

rater performance. *Assessment in Education: Principles, Policy and Practice*, *18*(3), 279–293. https://doi.org/10.1080/0969594X.2010.526585

Barth, P., & Stadtmann, G. (2019). Creativity ratings of fashion outfits presented on Instagram: Does gender matter? *Manuscript Sub-Mitted for Publication*. Frankfurt.

Barth, P., & Stadtmann, G. (2020). Creativity assessment over time: Examining the reliability of CAT ratings. *Journal of Creative Behavior*, *55*(2), 396–409.

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

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

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.

Cigularov, K. P., & Dillulio, P. (2020). Does rater job position matter in training needs assessment? A study of municipal employees in the USA. *International Journal of Training and Development*, *24*(4), 337–356.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. United States of America: Cengage Learning.

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

Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 1–19. https://doi.org/10.1177/0265532215582282

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Approaches to classroom assessment inventory: A new instrument to support teacher assessment literacy. *Educational Assessment*, *21*(4), 248–266. https://doi.org/10.1080/10627197.2016.1236677

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

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

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

Eraut, M. (2003). *Developing professional knowledge and competence*. Taylor & Francis e-Library.

Erman Aslanoglu, A., Karakaya, I., & Sata, M. (2020). Evaluation of university students' rating

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

behaviors in self and peer rating process via many facet Rasch model. *Eurasian Journal of Educational Research*, *89*, 25–46. https://doi.org/10.14689/ejer.2020.89.2

Esfandiari, R. (2019). How predictable ratings are : The role of personality traits. *Journal of Modern Research in English Language Studies*, *6*(3), 33–55.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79–102. https://doi.org/10.37546/jaltjj34.1-3

Foschi, M. (2000). Double standards for competence: theory and research. *Annual Review of Sociology*, *26*(1), 21–42. https://doi.org/10.1146/annurev.soc.26.1.21

Ganellen, R. J. (2007). Assessing normal and abnormal personality functioning: Strengths and weaknesses of self-report, observer, and performance-based methods. *Journal of Personality Assessment*, *89*(1), 30–40. https://doi.org/10.1080/00223890701356987

Gyagenda, I. S., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement*, *10*(3), 225–246.

Hariatul Hafidzah Mahmad Khory, Mohd Nazri Abdul Rahman, & Muhammad Azhar Zailaini. (2021). Pengurusan pentaksiran bilik darjah mata pelajaran bahasa arab berasaskan keperluan pembelajaran murid. *Jurnal Kepimpinan Pendidikan*, *8*(2), 41–57.

Jin, K. Y., & Eckes, T. (2021). Detecting differential rater functioning in severity and centrality: The dual DRF Facets model. *Educational and Psychological Measurement*, 1–25.

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.

Lam, R. (2019). Teacher assessment literacy: Surveying knowledge, conceptions and practices of classroom-based writing assessment in Hong Kong. *System*, *81*, 78–89. https://doi.org/10.1016/j.system.2019.01.006

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

Linacre, J. M. (1998). Rating, judges, and fairness. *Rasch Measurement Transactions*, *12*(2), 630–631.

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

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

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

McMillan, J. H. (2013). *Research on classroom assessment and research*. United States of America: SAGE Publications, Inc.

Muhammad Firdaus Mohd Noh, & Mohd Effendi Ewan Mohd Matore. (2020). Rating performance among raters of different experience through multi-facet rasch measurement (MFRM) model. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(2), 147–162.

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.

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

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

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

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.

Norshafinaz Abdul Sani, & Faridah Yunus. (2018). Amalan perancangan, pelaksanaan dan pentaksiran dalam proses pengajaran dan pembelajaran pranumerasi di tadika swasta. *Jurnal Pendidikan Malaysia*, *43*(2), 101–110. https://doi.org/10.17576/jpen-2018-43.02-10

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)*.

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

Panayides, P., & Tymms, P. (2013). Investigating whether aberrant response behaviour in classroom maths tests is a stable characteristic of students. *Assessment in Education: Principles, Policy and Practice*, *20*(3), 349–368.

Permana, N. S. (2017). Peningkatan mutu tenaga pendidik dengan kompetensi dan sertifikasi guru. *Studia Didaktika: Jurnal Ilmiah Bidang Pendidikan*, *11*(1), 1–8.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413–428. https://doi.org/10.1037/0033-2909.88.2.413

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

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

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.

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.

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

Stiggins, R. J. (1999). Evaluating classroom assessment training i n Teacher Education. *Educational Measurement: Issues and Practice*, *18*.

Volume: 7 Issues: 45 [June, 2022] pp. 141 - 155]
Journal of Islamic, Social, Economics and Development (JISED)
eISSN: 0128-1755
Journal website: www.jised.com
DOI: 10.55573/JISED.074512

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.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing*, *9*(3), 239–261. https://doi.org/10.1016/j.asw.2004.11.001

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluations. *The English Journal*, *74*(5), 49–55. https://doi.org/10.2307/817702

Uvie, O. M. (2021). Teachers' competency towards the implementation of school based assessment in secondary schools in Edo State, Nigeria. *International Journal of Education, Learning and Development*, *9*(2), 51–59.

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.

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

Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology*, *35*. https://doi.org/10.1111/j.1744-6570.1982.tb02208.x

White, A., Shen, F., & Smith, B. L. (2002). Assessing advertising creativity using the creative product semantic scale. *Journal of Advertising Research*, *36*(4), 241–243. https://doi.org/10.2501/JAR-41-6-27-34

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

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

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161–192. https://doi.org/10.1177/0265532216686999

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. (2018). *Pembinaan instrumen kompetensi guru pelatih dalam pentaksiran bilik darjah*. Universiti Kebangsaan Malaysia.

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.