Journal of Contemporary Issues in Business and Government Vol. 27, No. 2,2021
https://cibg.org.au/

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Rater Errors in Rating Processandthe Needsto Be Identified Among Student Raters

NurulFarienaAsli
Faculty of Education, UniversitiKebangsaan Malaysia
nurulfariena87@gmail.com (First author)

Mohd Effendi Ewan MohdMatore*
Faculty of Education, UniversitiKebangsaan Malaysia
effendi@ukm.edu.my(Corresponding author)

MelorMdYunos
Faculty of Education, UniversitiKebangsaan Malaysia
melor@ukm.edu.my

Abstract: In assessing performance-based language assessment, rater behaviour is one of the contributing factors in measurement error which is derived from rater error that become a threat in a rating process. The emphasis on students' self-directed learning after the implementation of CEFR in the Malaysian English language syllabus has required students to be able to assess their progress where the validity and reliability of the scores should be unquestionable. However, since it is a new practice, there is a lack of awareness of the need to identify rater behaviour among students. Therefore, this paper aims to discuss the different types of rater errors that occur in a rating process which will highlight the importance of the errors to be identified among students in secondary schools in Malaysia. Other than that, it is also aimed to propose a conceptual framework of rater-mediated assessment using the Many Facet Rasch Model (MFRM) that can be used in understanding the rater errors. The implication of this conceptual paper is teachers will gain insights into the factors that will become a threat for students to be good rater. Apart from that, the conceptual framework of rater mediated-assessment using MFRM will assist teachers to understand the relationship between three facets which are task, examinee, and rater with the outputs produced by MFRM. Future research should delve into factors that contribute to student's rater errors which undoubtedly affecting their judging in a rating process based on the conceptual framework of rater mediated assessment using MFRM.
Keywords: *rater errors, performance-based assessment, MFRM*

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Introduction

The landscape of language learning has undergone a few changes fromthe 1960's until the present mainly in the area of language testing and assessment due to the need to reflect on the importance of English as the main communicative language used (Mcnamara, 2014). This is where the traditional method of measuring students' performance is no longer become reliable and the focus is now on performance-based language assessment. It is believed that students' actual ability will be able to be measured on the skills that they are able to demonstrate rather than solely depends on the pen and paper examination. Realizing this, improvements have been made exclusively to the English language in the education system to suit that intent. This has been proven as more than 40 countries around the world including Malaysia have begun to adapt and implement the Common European Framework of Reference (CEFR) in their education system, an international framework for language learning (Council of Europe, 2020).

The CEFR has spread well beyond the boundaries of Europe in recent years and has become the primary guide in many contexts for language teaching (Panadero et al., 2018). To date, ongoing studies on how to use CEFR in the education system effectively in a particular local context are conducted around the globe for instance in Finland and Austria by (Holzknecht et al., 2018), United Kingdom (Díez-Bedmar & Byram, 2019; Green, 2018), Turkey (Özdemir-Yllmazer & Özkan, 2017), Canada (Figueras et al., 2013), China (Zeng & Fan, 2017; Zheng et al., 2016), Vietnam (Le, 2018; Nguyen & Hamid, 2020), Thailand (Foley 2019) and Malaysia (Alla Baksh et al., 2016; Ishak & Mohamad, 2018).

CEFR also emphasizes the role of learners to become independently responsible with their language learning as it provides the rating scale with 'can do' statements where learners will be able to assess their progress in learning a language (Council of Europe, 2001). The introduction of CEFR in Malaysia is a good effort done by the ministry in order to change the previous curriculum which does not yield expected results after learning the language in schools for eleven years. Among the real situations that occur are the lack of mastering language skills as well as the attitude of being spoon-fed in the learning process due to the exam-oriented system. Due to that, the English Language Roadmap prepared by the Ministry of Education in Malaysia has highlighted the need to train students to become independent learners (Zuraidah, 2015). Teachers are exposed to a variety of assessment strategies where self and peer assessment seems to be among the strategy that teachers need to use in the classroom. In order to train students to become raters and to be able to assess their progress and their peers' is a long process and cannot be achieved in a limited time. Scholars and researchers in the language testing and measurement field have come up with the term 'rater-mediated assessment' when the assessment is involving raters. Therefore, it must be bear in mind that students who conduct the self and peer assessment are also one of the raters in the category of rater-mediated assessment and have own behaviour in rating that might be different or the same with the expert raters' behaviour.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

What makes it crucial to measure the students' rater behaviour?

In Malaysia, the practice of making students a rater rarely happened as the system itself makes teachers the sole assessor for students' performance (Idris& Abdul Raof, 2017). However, with the new implemented English Language syllabus (CEFR aligned KSSM), students need to be responsible for their learning which requires the students to become reliable raters. Expanding out from Malaysia, the Southeast Asian countries are also struggling in training students to become raters since the system is also emphasizing teachers as the main assessor and teachers are also being skeptical to make students as the assessor (Foley, 2019). The main concern will be the rater errors that might occur will be a threat to the validity of the results gained and it must be controlled and minimized in some way (McNamara, 1996).A study on judging behaviour based on the rater errors was conducted in Malaysia by Abu Kassim(2011) which highlighted a few rater errors which are rater severity, restriction of range, central tendency, and internal consistency. Other studies were done by researchers (Ahmadi Shirazi, 2019; Eckes, 2012; Isbell, 2017; Myford & Wolfe, 2003, 2004; Wind & Engelhard, 2012; Wu & Tan, 2016) on rater errors also yield the same result which highlighting the same type of rater errors as studied by Abu Kassim (2011). Based on those studies, it has been shown that it is common to have the situation where rater errors occur in a rating process. Thus it emphasizes the importance of identifying the rater errors among students in a secondary school as it is a new practice in Malaysia new curriculum and teachers are still not aware of the rater errors' existence.

However, the variability associated with rater in a performance assessment is seen as extensive, difficult to control, and difficult to be eliminated (McNamara, 1996; Abu Kassim, 2011). In rater measurement, rater errors are known to be one of the contributing threats in a rating process. Thus, its existence in a rating process is crucial to be investigated in order to minimize its effect on the scores given by the rater. The behaviour of a rater is depending on what rater error occurs while giving scores. The issues of rater errors have been discussed thoroughly by (Myford & Wolfe, 2004) where it can be seen that some raters are prone to use the middle score in the ratings, dominantly use certain bands whether the highest or the lowest, give good scores to bad quality work vice versa, provide a high score for good students even though it is not a good piece of work and gives a low score to low ability students even though it supposes to gain a better score.All these judging behaviours are affecting the judgement of a rater and the scores given can be argued and questioned whether the result is reflecting on the students' actual performance in an assessment task. Once the result does not represent students' actual ability, it will affect their well-being in their future as they might not have the actual competency to be in a particular field. Therefore, it will lead to the worst consequence that might happen as the disqualified workers are able to get the position due to unfair judgement done by raters. Thus, it is crucial to measure the judging behaviour based on the rater errors in order to minimize its effect in a rating process.

Apart from that, studies on rater errors in assessing students' performance in language assessment found in the literature are mostly focusing on expert or novice raters among language teachers (Eckes, 2012), lecturers (Ahmadi Shirazi, 2019; Goodwin, 2016; Holzknecht et al.,

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

2018; Humphry & Heldsinger, 2019; Şahan & Razı, 2020; Wu & Tan, 2016), graduate L1 speakers in language teaching (Isbell, 2017) and language instructors (Polat, 2018). There is also a dearth of study aimingat school students as raters,especially in the Malaysian context. To date, there is only one study found which has involved students as raters by Idris and Abdul Raof(2017).However, this study focused on modest ESL pre-university students in assessing their performance as well as their peers in speaking skills.

After the alignment of CEFR with the English language syllabus, teachers need to train students to be able to assess their progress as well as their peers which requires them to become a rater. This has been mentioned in The English Language Roadmap which highlights that secondary school students are believed to have maturity in directing their learning (Zuraidah, 2015) therefore it is crucial to train the students to become good rater in order to assess their learning as well as their peers. Thus, it is a need to highlight the different types of rater error that might occur when teachers want to train their students in becoming a rater. The different types of rater will be a valid explanation on how students behave when they become a rater.

Literature Review

Rater Error

The existence of errors is the nature of most measurement factors, particularly raters have raised serious concerns regarding the psychometric quality of the scores awarded to examinees (Eckes, 2009). In writing assessment, the reliability of ratings has been an alarming issue for decades as there are always variations in the writing elements preferred by raters (Kayapinar, 2014). With the increasing implementation of performance assessment in second language writing where raters are involved using rating scales, the focus has been turned to the issues of raters when they do the rating process (Schaefer, 2008). Therefore, it is important to determine the quality of ratings obtained from raters by exploring the rater issues. Among the issues of raters, the one that has gained a lot of attention in the world of research is rater performance using different terms such as rater bias (Eckes,2009; Eckes, 2012; Schaefer, 2008), rater error (Abu Kassim, 2011) and rater effect (Ahmadi Shirazi, 2019; Myford & Wolfe, 2003).

Myford and Wolfe (2003) have discussed these terms used in the research literature paper and concluded that there is no clear distinction between these terms and has resulted to make the reader who is not used to the literature becoming confused. Due to that, this study will use the term rater error following the work of Abu Kassim (2011) as we are discussing rater-related errors in measuring performance. Rater errors that are explicated in this study are associated with rater severity, halo effect, central tendency, restriction of range, and internal consistency which are also being regarded as classic psychometric errors (Myford& Wolfe, 2003). Since the study on rater error is central to performance-based assessment, there is a need to explore students' rater error in order to determine their judging behaviour.

There are several studies found in the literature regarding rater-related performance as shown in table 1. It shows the different types of rater errors that they are focusing on. In addition to that, all studies are using teachers or lecturers as the raters, and studies on students becoming raters

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

are rarely found. It may be due to the fact that raters with experience or novice raters with knowledge in language assessment will create fewer rater errors and provide reliable and valid results for the studies. Realizing that school students also need to be trained to become raters in order to assess their learning, study on rater errors must involve students as well.

Table 1: Study on literature regarding types of rater

Num	Author	Title	Rater
1	Mohd Noh & Mohd Matore (2020)	Rating Performance among Raters of Different Experience Through Multi-Facet Rasch Measurement (MFRM) Model	Teacher
2	Şahan & Razı (2020)	Do experience and text quality matter for raters' decision-making behaviours?	Lecturer
3	Ahmadi Shirazi (2019)	For a Greater Good: Bias Analysis in Writing Assessment	Lecturer
4	Humphry & Heldsinger (2019)	Raters' perceptions of assessment criteria relevance	
5	Holzknecht et al. (2018)	Comparing the outcomes of two different approaches to CEFR-based rating of students' writing performances across two European countries	Lecturer
6	Polat (2018)	European Journal of Foreign Language Teaching Defining Severe Graders through MFRM.	English Language Instructors
7	Idris & Abdul Raof (2017)	The CEFR Rating Scale Functioning: An empirical study on Self and Peer assessment.	Student
8	Isbell (2017)	Assessing C2 writing ability on the certificate of English language proficiency: Rater and examinee age effects	Graduate L1 speakers in language teaching
9	Wu & Tan (2016)	Managing rater effects through the use of FACETS analysis: the case of a university placement test	Lecturer
10	Goodwin (2016)	A Many-Facet Rasch analysis comparing essay rater behaviour on an academic	Lecturer

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

		English reading/writing test used for two purposes	
11	Wang (2016)	Evaluating Rater Accuracy in Rater-Mediated Assessments Using an Unfolding Model	Lecturer
12	Eckes (2012)	Operational rater types in writing assessment: Linking rater cognition to rater behaviour	Language teachers
13	Wind & Engelhard (2012)	Examining rating quality in writing assessment: rater agreement, error, and accuracy.	Lecturer
14	Abu Kassim (2011)	Judging behaviour and rater errors: An application of the many-facet Rasch model	Lecturer

Types of Rater Error

Rater Severity

Among the different types of rater errors, rater severity is the most widely discussed and researched by scholars (Eckes, 2009; Eckes, 2012; Wang, 2016; Wu & Tan, 2016) because it is being referred to as the most serious error that a rater can produce during a rating process. Rater severity can be defined as the tendency for raters to award higher or lower ratings compared to what is possible to be justified by the performances (Engelhard, 1994). The reason for the occurrence of differences in rater severity is due to the fact that raters do not interpret the rating scale similarly as they have different standards and expectations towards writing. For instance, the same essay writing can be perceived as good, average, or poor by different raters. In order to determine the differences in rater severity, interrater-agreement or reliability will be examined as it portrays the extent to which raters agree in the ratings that they award.

Two studies are found in the literature within the Malaysian context. A study by Abu Kassim (2011) on rater severity involving university English Language instructors shows the effects of rater severity on the accuracy of the estimation of performance. The rater severity results need to be adjusted in order to get the actual performance estimation based on MFRM which indicate the occurrence of different severity level among the experienced raters. In addition to that, Mohd Noh & Mohd Matore (2020) in a study on rater performance among English teachers depicts that raters with different experiences exhibit non-uniform severity level whereas, the experienced raters displayed more consistency than the inexperienced raters. This is because experienced rater shows more understanding of the rubric and procedure of the assessment. Based on the above studies, it is a need to identify the severity level of the students when they become the rater in order to make sure the scores are accurate with their performance. In addition to that,

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

teachers must be aware that students are novice raters where some of them might have difficulties in assessing their progress due to the different levels of language proficiency. Therefore, in order to train students to become reliable rater, a study on their severity level will be crucial.

Halo Effect

Another detrimental rater error that contributes to the psychometric measurement in a rating process is the halo effect. It explains the raters' tendency to only focus on one aspect of a performance that will dominate their judgement when using a rubric (Wind & Engelhard, 2012). Meanwhile,Davis (2018) elucidates halo effect occurs when raters award similar scores across all categories in the rubric as they have difficulties in distinguishing the differences in aspects of performance. Based on that, it can be concluded that the halo effect happens when raters are using a holistic judgement in assessing performance even though they are using other types of rubric other than a holistic rubric. A typical halo effect example is when a rater awards the same score towards different aspects of performance (Abu Kassim, 2011). Thus, the halo effect happens when the rater has the tendency to give the same marks when they give scores. Raters who have halo effects do not have the ability to distinguish the differences in the construct from the rubric (Myford & Wolfe, 2004). It has been claimed that this type of error can be seen when analytic-type rating scales are used. An example of the halo effect is when the rater gives the same score for a different aspect of performance.

In the Malaysian context, having students become rater is a new practice where previously they depend on teachers to award them marks. In assessing performance-based language assessment, two essential elements involved are rater and rating scale (McNamara, 1996). Thus, undeniably the practice of giving the rating to their work will become a major challenge where students also need to be able to use the rating scale primarily as a rubric in order to assess their ability. A rubric is believed to bean instrument that can help raters to provide valid and reliable judgements in a rating process (Dickinson & Adams, 2017). Since scholars have highlighted that the halo effect will occur when raters are unable to distinguish the differences in the construct from the rubric, teachers need to provide students with a suitable rubric which only focuses on one aspect at a time. When the halo effect is identified among students, the next step to be taken by teachers is to study different types of the rubric and which rubric can be used effectively by their students.

Central tendency

Another type of rater error that will be the focus of this study is the central tendency. This type of rater effect will occur when the rater uses middle categories predominantly (Abu Kassim, 2011). It also exhibits the judging behaviour of the rater who is being reluctant to use extreme categories but overuses the middle category of a rating scale (Myford & Wolfe, 2003). It is a play-safe situation to avoid being too lenient or too strict and Abu Kassim (2011) mentions that it is a preference for a rater to only use somewhere around the middle categories in order for the ratings not to be far from those given by other raters.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

A remarkable study using the students in the pre-university course or known as form six students in a state in Malaysia has found that students who do the self-rating tend to award a moderate level of rating (Idris & Halim, 2015) because they are not confident in becoming a rater. Therefore, being in the middle categories will be the safest place for the students when they want to do the rating primarily when the rating is not involving their teachers' voice. They are afraid if they are not giving the scores as what their teachers might give to them. If this type of rater error occurs among students, the score given by students can be questionable. Once teachers have noticed that central tendency happens among their students, teachers need to boost students' motivation to be more confident in becoming a rater without the teacher's involvement.

Restriction of range

In contrast to the central tendency, restriction of range happens when ratings are limited to a few categories in the rating scales (Abu Kassim, 2011; Wind & Engelhard, 2012). The situation that might occur is when some raters will overuse the lower band of the scale meanwhile other raters overuse the upper band of the scale. A severe or lenient rater may be considered to portray this kind of rater error. It is also considered a serious threat to rating quality since the raters are unable to distinguish the differences in examinees' performance levels (Abu Kassim, 2011). The difference between central tendency and restriction of range is when the central tendency is used to describe the situation where the marks awarded are clustered in the middle range while restriction of range is where the ratings are clustered about any point on the rating continuum (Myford and Wolfe, 2004). It occurs when ratings are restricted to very few categories. Some raters will only give ratings using the lower end of the scale and others might give more ratings the upper end.

This type of rater error highlights that in a rating process, some raters are unable to differentiate the differences in examinees' performance level. This situation has raised the concern of this writer whether Malaysian students who are in the moderate or low level of proficiency are able to become valid and reliable raters as they have limited ability in the language.Azman (2016) has revealed that even though Malaysian students are taught English from the age of 6, it does not guarantee they have achieved competency in language acquisition. Therefore, there will be a problem when teachers want to train students to become rater who can assess their progress. By knowing that restriction of range does occur among students in a particular class, teachers are able to put a focus on a particular student and provide ample assistance in the area that students are not able to distinguish.

Internal Consistency

Most of the scholars also indicate another type of rater error that provides threats to the rating process which is internal consistency. This problem can be seen when raters are inconsistent in awarding marks of similar performances. Some raters award higher marks for poor performances and low marks for good performances due to any possible factors that may influence the raters such as exhaustion or carelessness. Linacre (1989) raised the concern on this type of rater error as it is more serious as the raters are inconsistent in their judgement compared to rater severity.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Internal consistency also shows that bias does occur since the raters consistently award low or high marks towards certain groups. Since the level of proficiency among students in Malaysia are varied, this error undeniably might occur due to the fact that students do not have the ability to rate especially students with low proficiency. When this type of rater error occurs, teachers are able to investigate why it happens to a particular student.

Bias

Another rater error that always occurs in a rating process is rater bias even though a scoring method or rubric is being used to award scores. Rater bias refers to a systematic pattern of rater behaviour that demonstrates itself in bizarrely severe (or lenient) ratings related to a particular aspect of the assessment condition (Eckes, 2012). Raters might exhibit more severity toward a specific group of examinees and is usually due to examinee level of proficiency or raters' preferences (Kondo-brown, 2002). In addition, bias also prone to happen on certain aspects that are being observed for example some raters rate grammar and vocabulary more harshly or leniently than other aspects (Wigglesworth, 1993). Apart from that, a study by Ahmadi Shirazi (2019) has found that raters' experience, L1, and educational background emerge as the sources of rater bias in a rating process. Thus, there are many factors that may contribute to rater bias which must be able to be controlled and minimized. However, no study can be found regarding rater bias in Malaysian context mainly involving school students as rater. Therefore, it is crucial for teachers to be able to identify rater bias among students and further investigations on the factors that influence bias among student rater in Malaysia need to be piloted. This will give insights on what are the main factors that influence rater bias among students and how it can be reduced. Figure 1 summarizes the types of rater that commonly occur in a rating process.

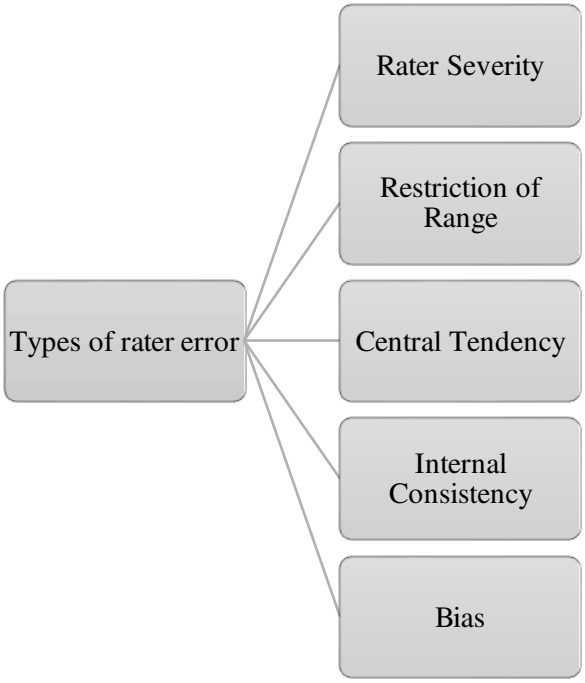

Figure 1. Types of rater error

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Recommendation

In this paper, it is proposed that the rater errors can be measured through Many Facet Rasch Model (MFRM). MFRM is particularly important in this aspect as it supports the observation and calibration of differences in rater severity, allowing these differences to be taken into account in the analysis of the rating assigned (Myford& Wolfe, 2004). In other words, MFRM does not require raters to score or assess in the same way as it identifies and tracks differences in raters' severity (Linacre, 1989).Many Facet Rasch Model is the extension of Rasch Measurement Model which is carried out using the computer program FACETS. In the late 1990s, the MFRM model was created to be used in mediating rater-effect in performance assessment such as speaking and writing (Lumley & Mcnamara, 1995). This program uses the scores given by the rater in the rating process to estimate individual examinee proficiencies, rater severities, and criterion difficulties, and scale category difficulties (Eckes, 2015). The results gained from MFRM is a graphical display showing the joint calibration of examinees, raters, criteria and the rating scale categories.

Previously, before the presence of MFRM, researchers have put the focus on the Classical Test Theory (CTT) or well known as the true score model. It is the earliest measurement theory that has been used for more than 80 years. The belief in CTT is it only measures anything that can produce a concrete quantitative number. In terms of the rating process, based on CTT, the use of multiple raters is functioned to control the variability as a result of rater errors. When more raters are involved in the process, the precision in measurement becomes increasing since there is more information available to estimate the performance. In addition to that, CTT also emphasizes that raters must agree in their judgement. The more similar the rating awarded to a task it will be resulted to the higher level of rater agreement and undeniably the inter-rater reliability is also higher. However, the assumption that raters should agree in their judgement is very vague to be supported. No two raters in their judgement of every performance they encounter can be perfectly unanimous (Engelhard, 1994; Linacre, 1989). Furthermore, Linacre explained that when raters know that their agreement in awarding marks is extremely important, they start to consider the other raters when assigning scores thus, it limits their independence in the rating process. This is where the constraint of the forced agreement has an influence on the raters (Abu Kassim, 2011).

Another measurement theory that will be discussed is the Generalizability Theory or G-Theory. The emphasis on sources of error has been extended and elaborated further based on the context of G theory as in contrast with Classical Test Theory (CTT). G-theory proposes that measurement errors arise from multiple sources rather than only a single undifferentiated measurement error (Brennan, 2011). In this theory, the potential sources of systematic measurement error are called facets, the levels of these facets are called conditions and the source of variance is normally the examinee proficiency (Eckes, 2015). Figure 2 below shows the decomposition of the observed score variance within G-theory by Eckes (2015).

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

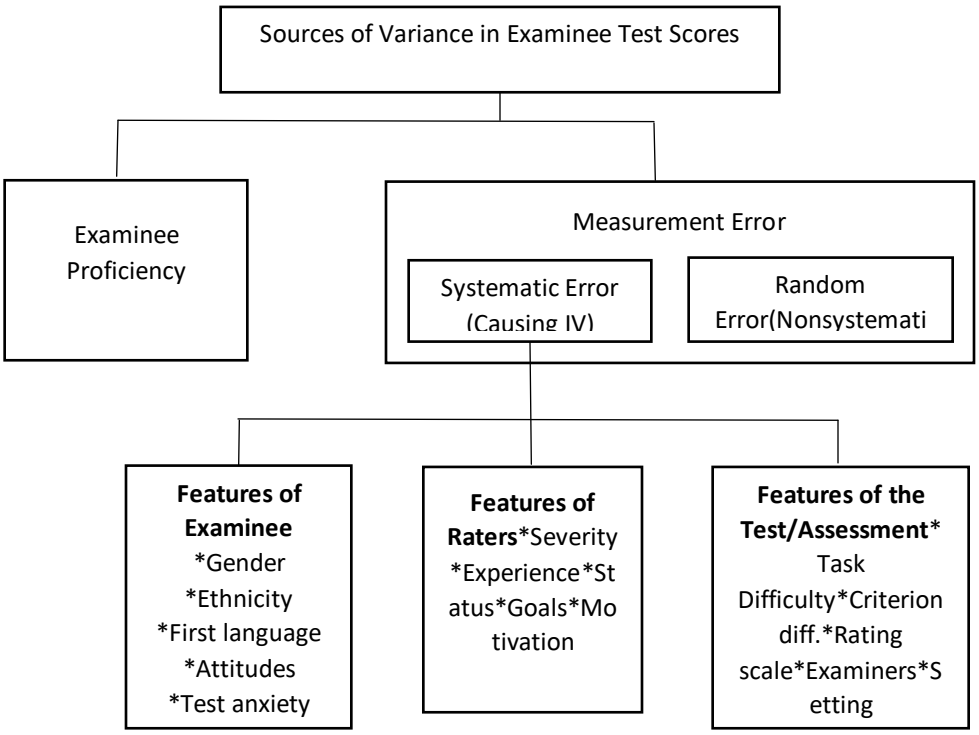

Figure 2.Decomposition of the observed score variance within G-Theory

The decomposition of sources of variance in examinee assessment scores is decomposed into true variance which is the examinee proficiency and error variance which are later decomposed into variance due to systematic error (construct-irrelevant variance; CIV) and random error. After that, the variance of systematic error is divided into three typical classes of potential sources of systematic error that represent the distal and proximal facets that are available in the MFRM framework. Due to that, it is one of the most similar theories to MFRM where it considers many of the same facets in its psychometric approach as MFRM.

Table 2. Differences between G-Theory and MFRM

G-Theory	MFRM
1. G-theory emphasizes rater homogeneity with a goal of making raters function interchangeably.	1. MFRM encourages rater self-consistency and expects raters to disagree with each other to some extent.
2. G-theory focuses on an examinee's total score as the unit of analysis and expresses this score in the ordinal metric of the original ratings.	2. The measures that result from a MFRM analysis (examinee proficiency measures, rater severity measures, etc.) have the properties of a linear, equal-interval scale if the data fit the model.
3. G-theory views the data largely from a	3. A MFRM analysis focuses more on

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

group-level perspective, disentangling the sources of measurement error and estimating their magnitude	individual-level information and thus promotes substantive investigation into the behaviour, or functioning, of each individual.

Apart from that, the G theory is also a well-known theory used by researchers in their studies about raters in SAPA (Ohta et al., 2018), however, in this particular study, G theory is not going to be used since there is another suitable measurement model that provides the intended result to achieve the objectives. This study is not aiming to research SAPA itself but rather use SAPA as the platform for a rater rating process. Since the aim of this study is to explore the rater errors among students and explain the judging behaviour of the students based on the rater errors, MFRM will provide specific information on what the researcher plan to do. The rater errors in the rating process can be modelled and statistically tested in MFRM as it is able to detect other rater errors such as restriction of range, halo effect, and internal inconsistency through the use of particular fit statistics (Myford& Wolfe, 2004; Linacre, 2014; Eckes, 2015; Engelhard & Wind, 2018).

The Conceptual-Psychometric Framework of Rater-Mediated Assessment in MFRM

The sample data analysis of the MFRM for rater error is based on a conceptual facet model by Eckes (2015) that can affect the performance assessment of the examinee to assist teachers and researchers in understanding the overall concepts of rater-mediated assessment. These aspects are outlined in Figure 3, illustrating some of the reciprocal relationships of rater errors and performance assessment, and comparing them to typical MFRM performance. However, in a specific assessment setting, the facets displayed do not reflect anything that can happen. Undoubtedly, the rating process is much more nuanced and dynamic than can be outlined in a diagram, and the facets that come into play at any given moment are diverse (Engelhard, 2018; Eckes, 2015).

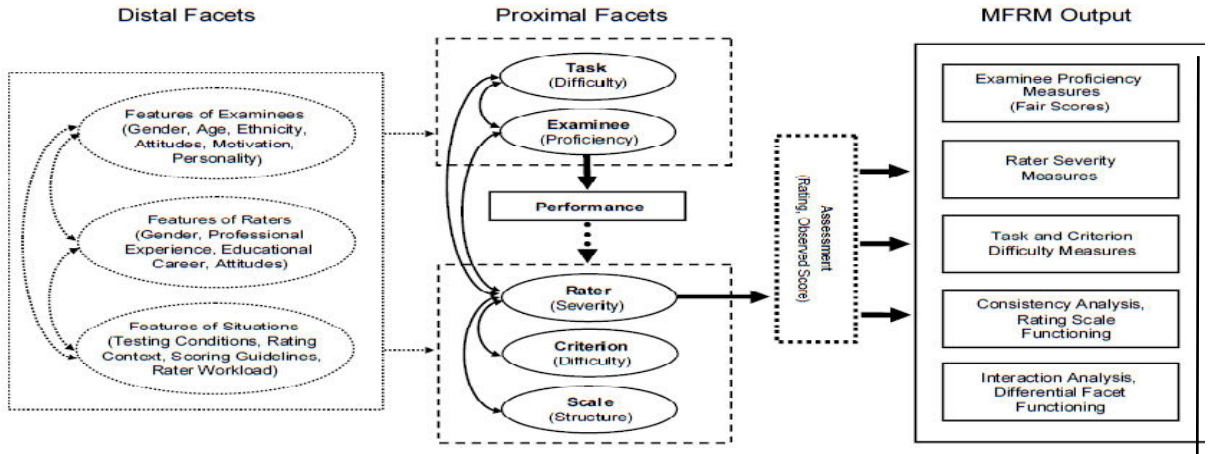

Figure 3. Psychometric Framework of Rater-Mediated Assessment

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Each facet is defined in an assessment and each interrelation of facets is a possible source of variance in the ranking. Assumptions about related facets may come from previous studies on the subject, from data obtained in similar contexts of assessment, or from earlier attempts at a simulation that have proven unsatisfactory. In the early stages of the measurement process, for example, some related facets can be invisible and these facets are called hidden facets. The existence of hidden facets may have adverse effects on the measurement performance, such as creating a biased assessment of the proficiency of the examinee. The structure refers to elements that are normally involved in a performance evaluation (Eckes, 2015).

In the middle part of Figure 3, the facets shown are called proximal facets. This segment includes facets that influence the scores awarded to the examinees immediately. The single most significant proximal factor is the ability of an examinee to represent the construct to be evaluated (e.g., writing skill). In an ellipse with an unbroken line, this facet is seen to illustrate the prominent role of examinee proficiency in the assessment process. The same graphic symbol is used for defining the other proximal facets. One of those other things relates to the task's difficulty or tasks to which examinees are asked to respond.In the prior knowledge brought to concentrate on a particular task, the examinees can vary considerably. This type of interaction, illustrated in the figure by a two-way arrow linking the examinee to the assignment, may have an effect on the examinee's output individually. Alternatively, it is likely that choosing a difficult task will result in a lower score than choosing a less challenging one because examinees can choose from a range of tasks that differ in difficulty. The overall effect will be an increase in the variability of the scores of the examinee, which indicates the result of an increase in the variation unrelated to the examinee.

Other proximal facets are unrelated to the construct in the lower section of the middle part and thus theoretically lead to systematic measurement error in the ratings that are rater severity, difficulty with the scoring criteria, and variability in the rating scale structure. The rating scale categories ordered can change their significance between raters, over time, between tasks, or between criteria (Weigle, 2010). In their interpretation of the ordering of scale categories, for instance, raters may vary from each other; that is, some raters may consider two adjacent categories or the performance levels indicated by those categories to be far closer to each other than others.

Figure 3 contains only two-way interactions, but there may be more than two-way interactions between proximal facets. For example, when evaluating low proficiency examinee, some raters appear to rank unexpectedly low scores on scoring criteria referring to task fulfilment, and unexpectedly high scores on scoring criteria referring to linguistic achievement when evaluating proficient examinee. A dotted arrow illustrates the relationship between success and the number of proximal facets related to the rater, indicating the difficulty of the rating process.

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Three kinds of variables that may have an additional influence on scores are shown on the left-hand side, but usually in a more indirect, guided, or diffuse way. Those variables are called distal facets. According to Eckes (2015), distal facets refer to (a) examiners' characteristics (e.g., gender, race, first language, personality traits, beliefs, goals), (b) raters' characteristics (e.g., number of foreign languages spoken, professional history, academic career, goals, and motivation), and (c) situational characteristics, that is, assessment or rating context characteristics (e.g., physical environment, rater workload, time of rating, paper-based vs.onscreen scoring, quality management policy). There might exist an interaction with proximal facets, or with the distal facets themselves. Therefore, that is why the relationship's complexity depends on a specific rating process context.

The overview of the major types of performance that can be obtained in the rater-mediated evaluation from MFRM analysis is on the right side of the framework. MFRM modelling thus offers a well-structured and in-depth description of the function played by each facet (proximal and/or distal) that is considered important in a given assessment context. Throughout the following, basic concepts are described in a non-technical way only to provide awareness of the variety of procedures available. An extension of the Rasch basic model is defined by the MFRM model. This extension is twofold: (a) it is not limited to just two facets (i.e. examinees and items) and (b) the data being examined is not dichotomous (Englehard& Wind, 2018). Therefore, more than one additional facet is taken into account in the analysis of performance assessment primarily raters, assignments, and criteria.

It is particularly when raters are interested in using the ordered scale categories (rating scales) as polytomous responses are included in the results. The key interests in certain assessment contexts apply to the examinees. A MFRM study defines a proficiency measure for each examinee (in logits). These measures accommodate between rater severities based on the concept of measurement invariance as the data fit the model. As such, for differences in the severity levels of the raters who participate in the rating process, the test proficiency measures are adjusted. In addition, the analysis generates standard errors that point out the precision of each measure of proficiency.

A separate parameter value that represents each facet is produced by this model. The first output, based on the framework, refers to a fair score reflecting the proficiency measures of the examinee. A fair score is an average score that is also known as the predicted score and it is for each of the examinees (Linacre, 2014). It can be obtained from a shift of estimates of examinee proficiency recorded in logits to the raw-score scale equivalent scores. The fair score of other examinees will be calculated by the score of a rater of average severity. From there, a fair score would demonstrate the impact of model-based compensation for variations in rater severity/leniency.

The second output, which is the severity of the rater, can therefore be obtained on this basis. If the MFRM is applied, the individual facets are simultaneously analyzed and calibrated to the

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

logit scale, which is a single linear scale. The joint facet calibration makes it possible to assess the severity of the rater using the same scale as the proficiency of the examinee, task complexity and difficulty of the criterion. A frame of reference that will be used to view the results is generated since all parameter estimates for the facets are put on a common scale. If the data indicates adequate model fit, the measurements of examinee competence, rater complexity, task difficulty and criterion difficulty can be directly compared.

Fit indices provide the estimation of consistency across examinees, activities, and parameters based on the rating made by the individual rater in terms of the rater facet. The consistency analysis for evaluating the functioning of the rating scale can therefore be obtained from MFRM. In addition, this consistency analysis focused on the inspection of rater fit indices plays a crucial role in deciding the rating behaviour of raters because, apart from severity/leniency, the fit indices can detect different rater effects, such as central pattern or halo effects (Myford& Wolfe 2003, 2004;). Generally, the input data of performance assessments that will be used in an MFRM analysis are ratings based on a set, or sets or ordered response categories.

Limitations

This paper applies an exploratory approach in synthesizing relevant and available literature to understand the issues of rater errors in a rating process. Due to the lack of studies done on students as the rater, the discussion about rater errors is formed based on the literature that describes rater errors as general. In addition, based on the analysis, the coverage of the literature is only revolved around the field of language assessment and it might yield different analysis results in other contexts.

Conclusion and Implication

In conclusion, it is pertinent for teachers to have knowledge regarding rater errors as well as on how to identify the judging behaviours of their students based on the rater errors. With this knowledge, teachers are able to train students to become reliable raters. Undeniably, this will help them become independent learners where they can assess their learning progress as well as their peers. Even though rater errors are prone to occur in a rating process, it is compulsory for the errors to be controlled and minimized in order to obtain valid and reliable results. In terms of theory, it is believed that the introduction of the conceptual framework of rater-mediated assessment using (MFRM) in determining rater errors will give more insights to understand the relationship between all the facets and outputs. Future research might want to put a focus on each of the facets based on students' different levels of proficiency consists of students with a high level of proficiency, intermediate level of proficiency, and low level of proficiency. It is also advisable for future research to delve into factors that contribute to student's rater errors which undoubtedly affecting their judging behaviours when assessing their work as well as their peers.

Acknowledgment

This work was supported by the Ministry of Higher Education (MOHE), Malaysia, and Faculty of Education, UniversitiKebangsaan Malaysia (UKM) through the Fundamental Research Grant

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Scheme (FRGS) under (Grant number: FRGS/1/2018/SSI09/UKM/02/1), and in part of Dana Penyelidikan FPEND (Grant number: GG-2019-034). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which helped them improve the content, quality, and presentation of this article.

References

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

Ahmadi Shirazi, M. (2019). For a greater good: bias analysis in writing assessment. *SAGE,9*(1): 1–14. http://journals.sagepub.com/doi/10.1177/2158244018822377.

Alla Baksh, M.A., Mohd Sallehhudin, A.A., Tayeb, Y.A. & Norhaslinda, H. (2016). Washback effect of school-based english language assessment: A case-study on students' perceptions. *Pertanika Journal of Social Sciences and Humanities,24*(3): 1087–1104.

Azman, H. (2016). Implementation and challenges of English language education reform in Malaysian primary schools. *3L: Language, Linguistics, Literature,22*(3): 65–78.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. *Cambridge University Press,* 1–264. https://rm.coe.int/1680459f97.

Davis, L. (2018). Analytic, holistic, and primary trait marking scales. *The TESOL Encyclopedia of English Language Teaching*.1–6. John Wiley & Sons, Inc.

Dickinson, P. & Adams, J. (2017). Values in evaluation – The use of rubrics. *Evaluation and Program Planning,65*: 113–116. https://dx.doi.org/10.1016/j.evalprogplan.2017.07.005.

Díez-Bedmar, M.B. & Byram, M. (2019). The current influence of the CEFR in secondary education: teachers' perceptions. *Language, Culture and Curriculum,32*(1): 1–15. https://doi.org/10.1080/07908318.2018.1493492.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Strasbourg, France: Council of Europe/Language Policy Division.

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

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Many-Faceted Rasch Model. *Journal of Educational Measurement, 31*(2), 93–112. https: //doi:10.1111/j.1745-3984.1994.tb00436.x

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

Figueras, N., Kaftandjieva, F. & Takala, S. (2013). Relating a reading comprehension test to the CEFR levels: A case of standard setting in practice with focus on judges and items. *Canadian Modern Language Review,69*(4): 359–385. https://doi/10.3138/cmlr.1723.359.

Foley, J. (2019). Issues on the initial impact of CEFR in Thailand and the region. *Indonesian Journal of Applied Linguistics,9*(2): 359–370.

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing,30*: 21–31. http://dx.doi.org/10.1016/j.asw.2016.07.004.

Green, A. (2018). Linking Tests of English for Academic Purposes to the CEFR: The Score User's Perspective. *Language Assessment Quarterly,15*(1): 59–74. https://doi.org/10.1080/15434303.2017.1350685.

Holzknecht, F., Huhta, A. & Lamprianou, I. (2018). Comparing the outcomes of two different approaches to CEFR-based rating of students' writing performances across two European countries. *Assessing Writing,37*(April 2017): 57–67. https://doi.org/10.1016/j.asw.2018.03.009.

Humphry, S. & Heldsinger, S. (2019). Raters ' perceptions of assessment criteria relevance. *Assessing Writing,41*: 1–13. https://doi.org/10.1016/j.asw.2019.04.002.

Idris, M. & Abdul Raof, A.H. (2016, November 25-27). *Modest ESL learners rating behavior during self and peer assessment practice*. Language Testing Forum, Reading, United Kingdom.

Isbell, D.R. (2017). Assessing C2 writing ability on the certificate of english language proficiency: Rater and examinee age effects. *Assessing Writing,34*(April): 37–49.

Ishak, W. I. W., & Mohamad, M. (2018). The implementation of Common European Framework of References (CEFR): What are the effects towards LINUS students' achievements? *Creative Education, 9*, 2714-2731. https://doi.org/10.4236/ce.2018.916205

Kayapinar, U. (2014). Measuring Essay Assessment: Intra-Rater and Inter-Rater Reliability. *Eurasian Journal of Educational Research,*(57): 113–135.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. Language Testing, 19(1), 3-31.

Le, H.T.T. (2018). Impacts of the Cefr-Aligned Learning Outcomes Implementation on Assessment Practice. *Hue University Journal of Science: Social Sciences and*

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Humanities,127(6B): 87.

Linacre, J. M. (1989). Many-facet Rasch measurement. MESA Press.

Linacre, J. M. (2014). Facets computer program for many-facet Rasch measurement, version 3.71.4. Beaverton, Oregon, Winsteps.com

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

McNamara, T.F. (1996). *Measuring Second Language Performance*.Pearson Education Limited.

McNamara, T.F.(2014). 30 Years on — Evolution or Revolution ? *Language Assessment Quarterly,11*(2): 226–232.

Mohd Noh, M.F. & Mohd Matore, M.E.E. (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): 1–16.

Myford, C.M. & Wolfe, E.W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement,4*(4): 368–422.

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.

Ohta, R., Plakans, L.M. & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales : A generalizability analysis. *Assessing Writing,38*(August): 21–36. https://doi.org/10.1016/j.asw.2018.08.001.

Özdemir-Yllmazer, M. & Özkan, Y. (2017). Speaking assessment perceptions and practices of English teachers at tertiary level in the Turkish context. *Language Learning in Higher Education,7*(2): 371–391.

Panadero, E., Broadbent, J., Boud, D. & Lodge, J.M. (2018). Using formative assessment to influence self- and co-regulated learning : the role of evaluative judgement. *European Journal of Psychology of Education* 1–43.

Polat, M. (2018). Defining sevre graders through Many Faceted Rasch Measurement. *European Journal of Foreign Language Teaching,3*(4): 186–198.

Şahan, Ö. &Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing,* 1–22.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*

Van Huy Nguyen & M. Obaidul Hamid (2020): The CEFR as a national language policy in Vietnam: insights from a sociogenetic analysis.*Journal of Multilingual and Multicultural Development*. https://doi.org/10.1080/01434632.2020.1715416.

Wang, J., Engelhard, G. & Wolfe, E.W. (2016). Evaluating Rater Accuracy in Rater-Mediated

P-ISSN: 2204-1990; E-ISSN: 1323-6903
DOI: 10.47750/cibg.2021.27.02.608

Assessments Using an Unfolding Model. *Educational and Psychological Measurement,76*(6): 1005–1025.

Weiqiang Wang (2016): Using rubrics in student self-assessment: student perceptions in the English as a foreign language writing context.*Assessment & Evaluation in Higher Education*.https://doi.org/10.1080/02602938.2016.1261993.

Weigle, S.C. (2010). *Scoring procedures for writing assessment*. (J. Charles Alderson & L. F. Bachman, Eds.)*Assessing Writing*. Cambridge: Cambridge University Press.(original work published 2002).

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

Wind, S.A. & Engelhard, G. (2012). Examining rating quality in Writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement*,*13*(4): 321–335.

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 Development35*(2): 380–394.

Zeng, Y. & Fan, T. (2017). Developing reading proficiency scales for EFL learners in China. *Language Testing in Asia,7*(8): 1–21. https://doi.org/10.1186/s40468-017-0039-y.

Zheng, Y., Zhang, Y. & Yan, Y. (2016). Investigating the practice of The Common European Framework of Reference for Languages (CEFR) outside Europe: a case study on the assessment of writing in English in China. *British Council*. University of Southampton.

Zuraidah, M.D. (2015). *English Language Education Reform in Malaysia: The Roadmap 2015-2025*. Ministry of Education Malaysia.