JTET

http://penerbit.uthm.edu.my/ojs/index.php/jtet
ISSN 2229-8932 e-ISSN 2600-7932

Journal of Technical Education and Training

The Effect of Rubric on Rater's Severity and Bias in TVET Laboratory Practice Assessment: Analysis using Many-Facet Rasch Measurement

Azmanirah Ab Rahman[1*], Nurfirdawati Muhamad Hanafi[2], Yusmarwati Yusof [3], Marina Ibrahim Mukhtar [4], Halizah Awang[5] Anizam Mohamed Yusof[6]

[1,2,3,4,5,6]Faculty of Technical and Vocational Education,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat Johor, 86100, MALAYSIA

*Corresponding Author

DOI: https://doi.org/10.30880/jtet.2020.12.01.006
Received 30th August 2018; Accepted 04th September 2018; Available online 31st March 2020

Abstract: Performance assessments such as laboratory practice in Technical and Vocational Education and Training (TVET) are difficult to measure because they are subjective in their characteristics, which can cause bias. The use of a rubric scoring scale is suitable for measuring students' practical competency as it could translate the qualitative criteria into quantitative forms on a grading scale. The purpose of this study is to analyse the effect of rater severity for the performance of component installation of an electronic circuit project using Many-Facet Rasch Measurement (MFRM). Nine raters have examined the competency performance of 68 students using a four-point rating scale rubric consisting of 16 items. This study was conducted at three vocational colleges. The results showed that all raters measured with different severities but very high consistency. The effect of rater severity did not exist at the group level but exhibited at the individual level. It can also be examined that all raters were biased on the item and also towards students, either by being severe or lenient. Thus, the rubric cannot eliminate the effect of bias during the assessment. However, it can help raters to be consistent so that they have the same understanding when translating the rubric. These findings have implications for the psychometric quality of the electronic circuit project performance assessment.

Keywords: *Performance assessment, rater severity, Many Facet Rasch Measurement*

1. Introduction

Performance assessment such as for laboratory practice in TVET education is a method of assessment that can measure students' abilities and competencies in real-world situations (Crooks, Kane, & Cohen, 1996; Gipps & Stobart, 2003; Waugh & Gronlund, 2013). However, performance assessments are exposed to validity and reliability threats (Gipps & Stobart, 2003). According to Arter and Mctighe (2001), the answers given by students in their assignments cannot be evaluated using machines because their answers are varied and subjective. According to Abdullah and Shukor (2012), performance tasks cannot be assessed properly if they are not categorised into specific behaviours and smaller features that can be evaluated easily, as well as observable. Performance assessments that are not properly planned and practiced fairly will cause disappointment among the students (Zaghloul, 2001).

*Corresponding author: azmanira@uthm.edu.my

Issues such as validity and reliability in performance assessment may affect and cause problems in measurement due to the biasness of the raters. There are many studies that examined the effects of raters on student performance such as rater tendencies for being severe or lenient in evaluations (Eckes, 2005; Farrokhi, Esfandiari, & Schaefer, 2012; Kondo-Brown, 2002; Lumley & McNamara, 1995; Lunz & Stahl, 1990; Schaefer, 2008), for the biasness of students (Farrokhi et al., 2012), and the tendency of raters to be biased on items (Lynch & McNamara, 1998; Wigglesworth, 1993). In addition, the experience of the raters (Barkaoui, 2008; Weigle, 1999), as well as the expertise of the raters could also contribute to measurement errors (Jones, 1999). Raters also tend to give different marks to the same students at different times, or two different raters tend to score differently on the same student (Nitko, 2004). Besides the raters, the rating scales also affect student scoring. The tendency of the raters to use the middle scale as a 'play it safe' strategy has led to the lack of diversity of abilities among students (Farrokhi, Esfandiari, & Dalili, 2011; Knoch, Read, & Von Randow, 2007; Kozaki, 2004; Myford & Wolfe, 2004). Raters are also exposed to using a scale either extremely severe or lenient along its continuum (Saal, Downey, & Lahey, 1980).

Effective assessments require fair criteria and procedures in order to reduce bias. Therefore, the use of rubrics could be an alternative way to handle the issue (Allen & Tanner, 2006; Reddy & Andrade, 2010). A rubric is a systematic method for evaluating the quality of students' work based on different performance criteria. It is a great scoring tool for measuring authentic tasks that assess students' work and work processes (Montgomery, 2002). The rubric features are designed to provide an explicit scoring standard to users (Mabry, 2004). With the use of the rubric, the effect of bias can perhaps be reduced.

Most of the learning in TVET education is a performance-based learning such as practical skills. Practical learning is essential to support lectures in order to enhance students' understanding of theoretical concepts. For the Electronic Technology programme at vocational colleges, students will perform hands-on activities in laboratories. This course exposes students to the circuit development process on printed circuit boards (PCB) comprising four domains, namely circuit design process, etching process, component installation process, and troubleshooting process. Assessments of work process and product in this subject are difficult because they involve many criteria that need to be translated into the rubric form so that they are easier to evaluate and also reduce the effect of biasness. Therefore, this research is carried out to examine the rater severity effect when using the rubric of practical assessment so that the results measure the students' actual ability.

1.1 Rater Severity Effect

The effects of raters often discussed are the effects of severity and leniency, halo effects, and central tendencies. However, this article only discusses on the effects of severity. The conceptual definition of rater severity by DeCotiis (1977) is:

> "The set response is generally associated with a simple rating provider (positive loops) in one situation and a "hard nose" rating (negative loops) on the other. We can expect the severity to be expressed according to extreme high or low ratings with a slight diversity.

Meanwhile, Borman (1977) gave a slightly different definition of rater severity, also known as the "rating effect". On the other hand, the definition by Saal and Landy (1977) on the effect of severity and leniency of raters is "the tendency to give a higher or lower rating to the individual is more than permitted by the individual's behaviour". Engelhard (2013) also gave a similar definition; "the tendency of the raters to score high or low compared to student performance".

The definitions given by these four researchers are consistent in which that the raters give too high scores to low-skill students or give too low scores to high-achieving students, but in fact they provide a difference in meaning, conceptually and operationally. In the context of the Many-Facet Rasch Measurement (MFRM), the raters are said to be severe when they are likely to give lower scores than expected, if the scores provided by other raters are within the acceptable scores of the same student. Meanwhile, lenient raters tend to give higher marks than expected if other raters' scores are within the acceptable scores of the same student (Myford & Wolfe, 2004).There are various factors that contribute to the tendency of the raters to either be severe or lenient. Among them are professional experiences, personality traits, attitudes, demographic characteristics, workloads, and types of tasks. Raters who are senior or more experienced will be more severe in scoring. The senior raters will set the standard for other raters, taking into account their lack of performance that might be overlooked by other raters. Raters with less experience will benefit the students, especially when the students' performance is at the border of the total score (Eckes, n.d.).

Based on the definition of rater severity by MFRM five objectives have been determined. The objectives of this study are to: (i) determine the relationship between students' competence, raters severity, and item difficulty on one logit scale, (ii) determine the severity and consistency of the raters in assessments, (iii) examine the effect of severity and leniency of raters at the group level and individual level, (iv) investigate significant effects of rater-item bias, and (v) investigate significant effects of rater-student bias.

2. Methodology

This study was conducted by using the exploratory mixed method approach with the aim to examine the rater severity effect by using rubric in TVET for Electronic Technology course. Students should complete four main tasks which are design circuit, etching process, component installation process and testing and troubleshooting. These articles focus on component installation process only. The judgement process was difficult because it is subjective in their characteristics which can cause rater bias. The use of a rubric scoring scale is suitable for measuring students' practical competency as it could translate the qualitative criteria into quantitative forms on a grading scale.

A mixed method was carried out with the qualitative data collection and followed by the quantitative data collection. The qualitative data collection was used to construct the criteria and items for the rubric, consisting of content analysis of documents, interviews, and field work. These criteria and items obtained were used to develop the rubric. Then, the developed rubric was validated by seven content experts in the electronic engineering discipline. The validity and reliability of the rubric were analysed using a Rasch model through Winstep programme 3.68.2. The separation index obtained was 3.82 and this value indicated the separation of item difficulty. Meanwhile, the value of item reliability was 0.94. The value for reliability index was greater than 0.8 was considered as good.

The quantitative data approach was carried out to examine the rater effect severity. Nine raters from three vocational colleges were purposely selected to examine the competency of 68 students in performing the component installation of an electronic circuit. Three raters from each college were involved in the study. All raters were those who were teaching in the field of electronics with over 10 years of experience and able to recognise all of the students. This was to help smoother the process as the evaluation involved not only the final product but also the attitude and the work process. Table 1 shows the assessment plan that was designed before the assessment. A good assessment plan is when all raters could assess all students' work at once (Linacre, 1994). However, the assessment is still feasible even if the network is reduced, as long as students, raters, and items are connected to each other.

Table 1 – Assessment Plan

Vocational College	No. of Students	Raters								
		1	2	3	4	5	6	7	8	9
A	1-24	x	x	x						
B	25-50				x	x	x			
C	51-68							x	x	x

Before the evaluation process, the raters were briefed on the rubric as well as the exercises on the rubric. The purpose of the briefing was to provide an overview of the study, biased sources in the scoring, and description of the rubric. During evaluation, the researcher was not present with the raters. This was to make the raters more comfortable and could avoid scoring as 'play-it-safe strategy'. After all the processes were completed, the data were analysed using the FACET programme version 3.71.4 by MFRM.

MFRM is a continuation of the RASCH measurement model (Wright & Masters, 1982). This model can be used to measure more than two interacting facets to produce an array of facet observations such as rater severity, item difficulties, and student abilities on the same single logit scale. The severity of the raters can be directly analysed from the MFRM on the behavior of the raters by analyzing the data obtained (Eckes, 2011). According to Myford and Wolfe (2004), there are two ways to detect the severity and leniency of the raters, which are at the group level and at the individual level.

The group analysis of rater severity can be identified by the fixed chi square test, the rater separation ratio, the rater separation index, and the rater reliability index. The non-significant homogeinety index indicate the effect of leniency and severity. The raters should share the same severity after considering the measurement errors. The rater separation index means the statistically significant difference in the severity of raters within the raters, including in the analysis. Specifically, this index illustrates the real difference in the variance error unit. The reliability index of the raters is an indicator that shows how far the raters are separated to determine their facets. Reliability index should be close to zero, which indicates that the raters have the same level of severity (Myford & Wolfe, 2004).

3. Research Findings

3.1 The relationship between students' competence, severity of the raters, and difficulty of the item

Student ability, rater severity, and item difficulty were analysed at one common log odds on the variable map as shown in Figure 1. The first column is a logit scale, while the second column on the variable map illustrates the students' position according to their ability. An asterisk (*) represents two students. Students located at the very top are students with high abilities, while the lowest ranked students are those who were less capable. A total of 97% of students have

the ability which exceed the mean value (0.00). Only two students are below the mean of 0.00. However, there are 41 students (60%) who have high abilities so that no item is difficult to measure the abilities of the students.

The third column on the variable map illustrates the severity of the raters. A total of nine raters examined the competencies for the domain of component installation of an electronic circuit. Amongst all the raters, 4 of them have a severity of more than 0.00 and five raters have below the mean of 0.00. The top raters are the most severe raters, while the bottoms are the most lenient raters.

The fourth column on the variable map is an item difficulty. There are 16 items of the component installation of an electronic circuit domain. From the diagram, it can be seen that the most difficult item is the existing knowledge, while the easiest item is about safety. Of the total items, 6 items were above the logit 0.00 and 10 were below the 0.00 logit.

3.2 Rater Severity and Consistency

Measurements of the severity and consistency of the rater can be referred in Table 2. Based on the findings, the lenient rater was rater 5 with a logit value of -1.06 (SE = 0.1), while the most severe rater was rater 4 with a 1.01 logit value (SE = 0.08). The difference between the lenient rater and the severe rater was 2.07. The standard error (SE) for raters was between 0.08 and 0.1. This showed that the accuracy of measurement was high, where the SE approaching zero was better.

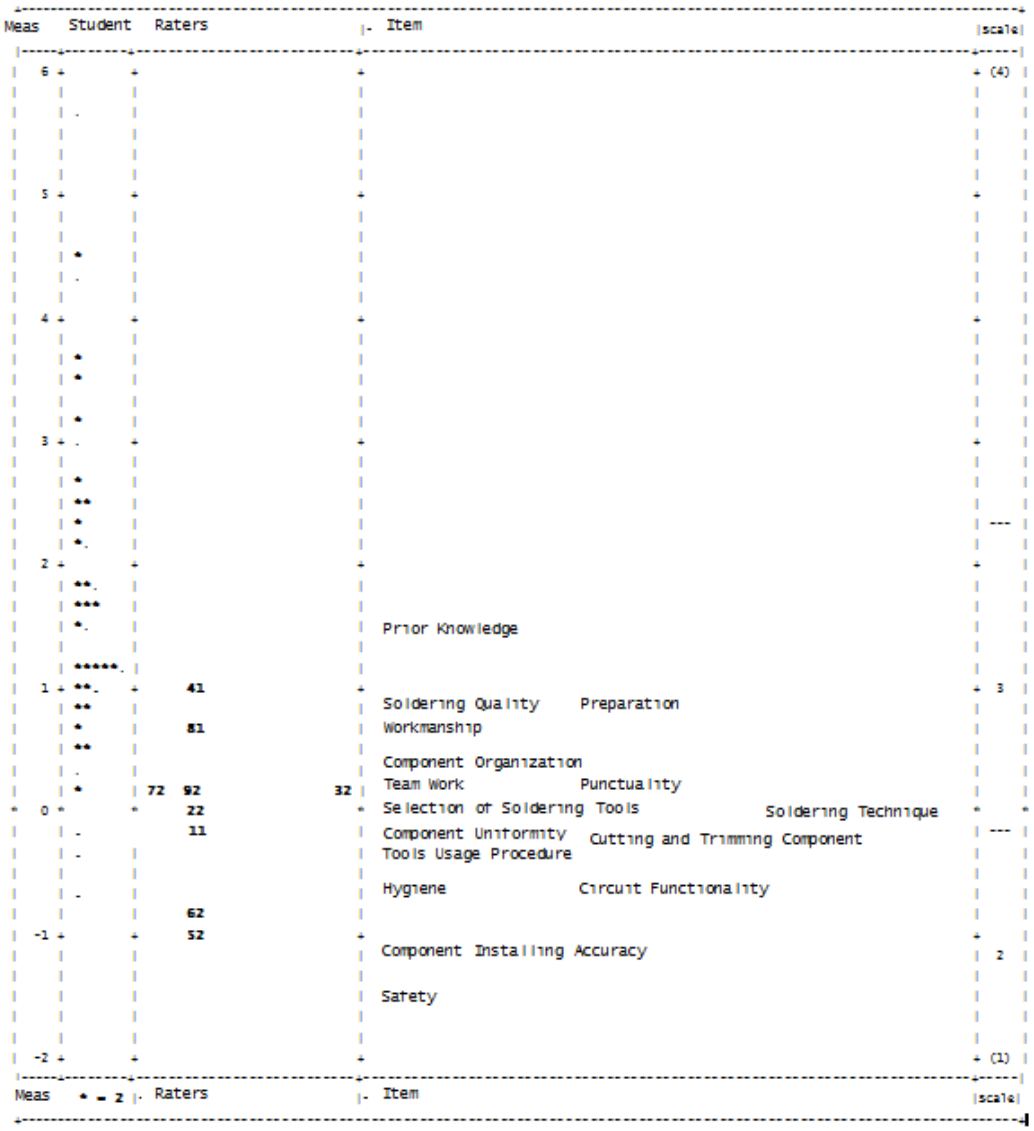

Fig. 1- Variable map for students' ability, raters severity and items difficulty of component installation for the Electronic Circuit Project

Consistency of the rater was also examined with the infit mean-square (MNSQ). Infit is more important than outfit in judging the rater fit (Myford & Wolfe, 2003). Infit and outfit mean squared have a range value of 0.5 and 1.5, respectively (Linacre, 2002). The fit range between these values is productive for measurement (Linacre, 2002). Infit

MNSQ values less than 0.5 are too consistent and have too little of variations, while values larger than 1.5 are too inconsistent and have too many variations (Myford & Wolfe, 2003). Based on Table 2, the infit MNSQ value ranges from 0.83 to 1.18 and the value of outfit MNSQ is 0.88 to 1.33. This value is within the allowable range.

The separation statistics were applied to analyse rater severity, namely, the homogeneous index, separation index, and reliability index as shown in table 3. The first statistic was the rater homogeneous indices. The rater homogeneous indices showed significant differences between raters where χ^2 (8, N = 9) = 395.0, p <0.05. This means the raters assessed with different severities. The analysis of the separation index was to show the statistical difference of the level of rater severity. The separation index for the rater was 7.07. This showed that there were seven categories of raters who provided different scales of the score. The index for reliability was between 0 and 1. The value approaching 1 indicated high reliability. However, the low reliability index was required by the most desirable result to have a reliability of rater separation close to zero, which would suggest that the raters were interchangeable and exercised very similar levels of severity (Myford & Wolfe, 2003). From the analysis obtained, the reliability of the raters was very high, which was 0.95. The high degree of rater separation reliability implied that the raters were differentiated in terms of the levels of severity they exercised. There is some evidence of unwanted variations between raters in their levels of severity (Myford & Wolfe, 2003).

Table 2 - Measurement Results for Rater Severity

Rater	Severity Measure (logits)	Standard Error (SE)	Infit MNSQ	Outfit MNSQ
5	-1.06	0.10	0.83	0.89
6	-0.87	0.1	1.04	1.33
1	-0.19	0.09	1.11	1.06
2	-0.04	0.09	1.18	1.10
9	0.09	0.09	0.91	0.92
3	0.17	0.09	0.99	0.96
7	0.22	0.09	0.86	0.88
8	0.67	0.09	1.07	1.08
4	1.01	0.08	0.87	0.88

Note: separation ratio index = 7.07; separation index = 9.76; reliability index = 0.98; Chi square value: 395.0; df: 8; significant (probability): 0.00

3.3 Effect of Rater Severity and Leniency at the Group and Individual Levels

Rater severity can be traced using two methods: at the group level and individual level. At the group level, four separation statistics were used to detect the rater severity and leniency, which are: (i) the Chi-Square test, (ii) separation ratio (iii), separation index, and (iv) reliability index. Table 3 refers to the measurement of severity and leniency at the group level.

Table 3 - Separation Index Statistics

Statistic	
Chi-Square test	395.0
Raters separation Ratio	7.07
Indexes raters separation	9.76
Indexes reliability	0.98

Based on the findings, the chi-square value of 395 with 8 degrees of freedom was statistically significant ($p < .005$), signifying that the raters did not all exercise the same level of severity when evaluating students. This indicated that there was a difference in the severity between the raters while scoring the students using the developed rubric. Rater separation ratio was 7.07, which showed that the difference between the rater severity was seven times greater than the error of the measured severity. The rater separation index was 9.76 (approx. 10), which showed that there were ten differences in the strata of the rater severity in the sample. The reliability index of the raters was 0.97, which clearly showed that the raters had varying severities. The significant chi-square values, high rater separation ratios, high rater separation indexes, and high reliability index indicated that no severity effect at the group level was detected.

To identify rater severity at the individual level, Table 4 is referred to. The larger the measure, the more severe the rater, whereas the smaller the measure, the more lenient the rater. The standard error for each severity measure appears in the Standard error column (SE). Standard error for raters was between 0.08 and 0.1, which showed that the accuracy of measurement was high. Based on the findings, seven raters out of total had a severity of measurements approaching the mean value 0.00. The range of the seven raters was between -0.87 and 0.67. There was a severe rater (Rater 4) and a lenient rater (Rater 5). The value of the severity measurement for Rater 4 (logit = 1.01) was a conspicuous outlier, which was 13 points above the standard error of the mean severity within the group. The value of the severity

measurement for Rater 5 (logit = -1.06) was 10 points above the standard error under the mean severity within the group. This finding showed that individual severity existed in the group.

Myford and Wolfe (2003) suggested that raters tended to use the rating scale in a different manner than other raters. Fair average might be used to pinpoint the rater who tended to use the rating scale differently. The fifth column shows the average (M) fair. The fair-M average indicates the observed average deviation of the raters from the total mean of the raters (Myford & Wolfe, 2003). The fair average for the most severe raters was 2.9, while the fair average for the lenient rater was 3.61. The difference in average among the most severe raters was 0.71 points lower than most lenient raters. On average, the ratings of Rater 5 tended to be almost one rating scale category higher than the ratings of Rater 4.

Table 4 - Measurement of Rater Severity and Leniency Individually

Rater	Severity measure (logits)	Standard Error (SE)	Observed score	Average (M) Fair
5	-1.06	0.10	3.56	3.61
6	-0.87	0.10	3.52	3.55
1	-0.19	0.09	3.26	3.34
2	-0.04	0.09	3.20	3.29
9	0.09	0.09	3.09	3.24
3	0.17	0.09	3.13	3.21
7	0.22	0.09	3.04	3.20
8	0.67	0.09	2.86	3.03
4	1.01	0.08	2.92	2.90

3.4 Raters-items Bias Interaction

The effect of the raters was also analysed on the item. There were 144 raters-items bias interaction. Of the total, there were 64 significant bias interactions. Lenient raters on the item have given a higher observation score than the expected value, with a *t* value greater than +2. Severe raters on the item have given a lower observation score than the expected value, with a *t* value lower than -2. From the analysis carried out, all raters contributed to significant bias interactions, either severe or lenient. Figure 2 is a summary of the analysis that has been described through the histogram. Based on Figure 2, Rater 7 is the most lenient towards a specific item of 6 interactions. Raters 1, 2, and 3 were also lenient to certain items of 4 interactions, whereas Raters 4, 5, and 9 were less or equal to (≤) 2 severe bias interactions. Raters 1 and 3 contributed most to the leniency effects of the item, followed by Raters 2 (5), 7, and 9 (4 respectively), and Raters 4, 5, and 8 (2 respectively). This finding showed that the raters had a high severity / leniency effect on the items. However, the effect of severity was higher (53%) than the effect of leniency (46%). Figure 2 displays the details of the findings.

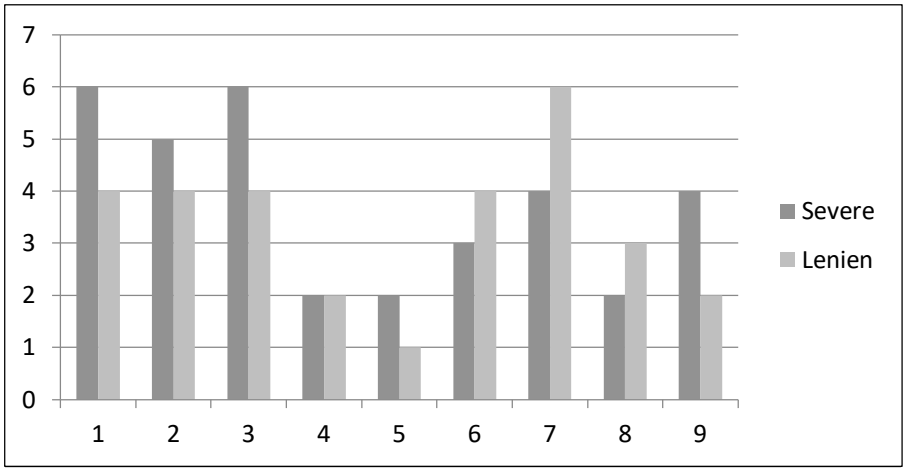

Fig. 2 - The distribution of rater analysis

Table 5 shows the total of bias interactions in which 34 bias interactions showed severe raters and 30 bias interactions indicated lenient raters. An example of a lenient interaction was Rater 8, which had a big difference between observation and expectation of 6 points (48-41.71 = 6.3), whereas the item was the most difficult item. At the

same time, Rater 8 was also severe on the simple items as students scored much lower than the expected score of 17 points (39 - 55.92 = -16.92). The findings showed that there was rater severity affecting certain items.

Table 5 - Rater-Item Interaction Report

Rater	Category	Observation	Expectation	Error	t-score	Infit MNSQ
1	*Selection of Soldering Tools*	96	78.7	1.64	2.99	0
1	*Soldering Technique*	95	79.74	1.03	3.67	0.9
1	*Component Uniformity*	91	79.28	0.51	3.89	0.9
1	*Cutting and Trimming Component*	91	79.97	0.51	3.71	0.9
1	*Workmanship*	65	72.37	0.31	-2.45	0.7
1	*Preparation*	62	70.99	0.3	-2.93	0.6
1	*Circuit functionality*	77	83.8	0.35	-2.61	0.8
1	*Prior Knowledge*	54	65.19	0.3	-3.42	0.8
1	*Tools Application*	72	80.67	0.33	-3.18	0.4
1	*Punctuality*	66	76.62	0.31	-3.7	0.3
2	*Selection of Soldering Tool*	95	77.45	1.03	3.96	1.1
2	*Soldering Technique*	95	78.53	1.03	3.83	1.1
2	*Component Uniformity*	87	78.05	0.42	3.1	1.2
2	*Cutting and Trimming Component*	87	78.77	0.42	2.89	1.2
2	*Prior Knowledge*	53	63.54	0.3	-3.18	0.9
2	*Hygiene*	74	81.88	0.34	-2.94	0.5
2	*Punctuality*	66	75.31	0.31	-3.19	0.6
2	*Preparation*	58	69.51	0.3	-3.67	1
2	*Tools Usage Procedure*	70	79.48	0.32	-3.42	0.6
3	*Selection of Soldering Tools*	92	75.76	0.56	4.87	1.3
3	*Soldering Technique*	87	76.88	0.42	3.45	1.2
3	*Uniformity*	85	76.38	0.4	2.96	0.9
3	*Cutting and Trimming Component*	83	77.13	0.38	2.05	0.9
3	*Organisation*	79	84.19	0.36	-2.01	0.6
3	*Prior Knowledge*	53	61.32	0.3	-2.49	0.7
3	*Tools Usage Procedure*	71	77.87	0.33	-2.43	0.4
3	*Safety*	77	82.33	0.37	-2.19	0.9
3	*Hygiene*	73	80.38	0.34	-2.7	0.3
3	*Punctuality*	62	73.54	0.3	-3.86	0.5
4	*Tools Usage Procedure*	88	76.05	0.4	3.82	1.2
4	*Punctuality*	81	71.05	0.35	3.1	2.4
4	*Cutting and Trimming Component*	69	75.18	0.31	-2.01	0.6
4	*Teamwork*	65	72.46	0.31	-2.37	0.7
5	*Soldering Technique*	96	90.39	0.59	2.33	1.1
5	*Teamwork*	84	88.96	0.37	-2.03	0.4
5	*Selection of Soldering Tools*	83	89.64	0.36	-2.76	1
6	*Tools Usage Procedure*	99	90.02	1.05	2.97	1.2
6	*Hygiene*	98	91.71	0.77	2.58	1.2
6	*Preparation*	90	82.22	0.43	2.72	1
6	*Prior Knowledge*	85	77.2	0.37	2.57	1
6	*Uniformity*	84	88.97	0.37	-2.04	0.7
6	*Teamwork*	82	87.81	0.35	-2.3	0.4
6	*Selection of Soldering Tools*	76	88.53	0.33	-4.95	0.9
7	*Hygiene*	75	62.2	1.03	3.45	1

Table 5 – (Continue)

Rater	Category	Observation	Expectation	Error	t-score	Infit MNSQ
7	*Component Organization*	74	65.51	0.75	3.01	1
7	*Circuit Functionality*	70	62.96	0.49	2.7	0.8
7	*Safety*	72	67.18	0.56	2.01	0.9
7	*Workmanship*	60	52.44	0.38	2.63	0.3
7	*Punctuality*	62	56.33	0.39	2.07	0.5
7	*Cutting and Trimming Component*	53	59.4	0.35	-2.43	0.4
7	*Soldering Technique*	43	51.08	0.33	-2.74	0.2
7	*Selection of Soldering Tools*	43	58.23	0.33	-5.55	0.2
7	*Soldering Technique*	40	59.19	0.33	-6.93	0.2
8	*Punctuality*	62	52.82	0.39	3.21	0.6
8	*Preparation*	55	47.23	0.36	2.56	0.6
8	*Prior Knowledge*	48	41.71	0.33	2.03	0.4
8	*Selection of Soldering Tools*	45	54.89	0.33	-3.5	1.9
8	*Soldering Technique*	39	55.92	0.33	-5.86	1.1
9	*Circuit Functionality*	71	63.78	0.52	2.77	1.1
9	*Prior Knowledge*	54	47.13	0.35	2.26	0.4
9	*Uniformity*	54	59.68	0.35	-2.16	0.5
9	*Soldering Quality*	44	52.18	0.33	-2.81	0.3
9	*Cutting and Trimming Component*	53	60.3	0.35	-2.8	0.3
9	*Soldering Technique*	47	60.09	0.33	-4.97	0.6

3.5 Rater-Student Bias Interactions

An analysis of rater–student bias interactions is displayed in Table 6. Of the total bias interactions (204), there were 14 significant bias interactions. The t-score value above +2.0 indicated that the raters were consistently scoring more leniently for some students. On the other hand, the t-score value below -2.0 indicated that raters consistently scored more severely for other students. Four of the students were assessed leniently by four raters, while 10 others were severely rated by 4 raters. Examples of lenient raters occurred when the marks given to student 26 were 54, but the expectation of the Rasch model were 47. Hence, the difference in the score was 7 points (54 - 47 = 7). Examples of severe rater occurred when the marks given to student 44 were 31, whereas the expectation of the Rasch model was 37. Thus, the difference in the score was 6 points (31-37 = -6). There were also severe raters for certain students but they were assessed leniently by other raters. Student 47 was rated severely by Rater 5 but examined leniently by Rater 6. Likewise, student 54 was examined leniently by Rater 8 but examined severely by Rater 9. The findings showed that there was a rater severity effect on certain students.

Table 6 - Rater-Student Bias Report

Rater	Students Ability (Logit)	Observation Score	Expectation Score	Error	t – score	Infit MNSQ	Rater
	1.87	54	47.00	0.44	2.74	0.8	4
44	0.47	31	36.97	0.36	-2.12	0.7	4
47	-0.61	33	44.29	0.36	-4.17	0.5	5
47	-0.61	54	42.96	0.44	4.09	0.9	6
40	0.74	57	51.47	0.49	2.29	1.0	6
26	1.87	52	57.11	0.43	-2.45	0.4	6
33	3.13	57	61.18	0.49	-2.65	0.7	6
41	3.75	59	62.32	0.54	-2.57	0.7	6
35	4.44	61	63.1	0.65	-2.14	1.1	6
54	0.5	47	39.84	0.39	2.59	0.2	8
60	1.56	38	47.24	0.36	-3.61	1.8	8
56	1.62	58	51.08	0.51	2.81	1.1	9
60	1.56	56	50.76	0.47	2.16	0.7	9
54	0.50	34	44.04	0.36	-3.73	0.8	9

4. Findings and Discussions

The MFRM analysis of the component installation of the electronic circuit project performance assessment rubric revealed interesting findings about: (i) the relationship between students' competence, the severity of the raters, and the difficulty of the item, (ii) rater severity and consistency, (iii) rater severity at the group and individual levels, (iv) rater–item bias interaction, and (iv) rater-student bias interaction.

Overall, the students had a high level of competence compared to the severity of the raters and item difficulties. Over 80% of students had a high level of competence. This finding is consistent with the findings of Iramaneerat & Yudkowsky (2006) and Matsuno (2009). According to Matsuno (2009), two things that might happen are considered; students really have a high level of ability or raters tend to give a high score to students with low abilities. Based on the findings, some assumptions can be made such as the raters could not differentiate the rating scale properly, the raters were influenced by student attitudes, and the number of ratees was too many that the raters gave scores by guessing. Raters tend to follow the previous criteria or they were less familiar with the content of the descriptor of the rubric. The difference in the level of severity will affect the student's competence. Highly competent students are influenced by the scores provided by the raters while the difficult items also influenced by the scores given by the raters. Although the scale of the rating is established, it depends on the raters who are using it.

The rater severity and consistency of this study clearly showed that raters measured with different severity but had high consistency. This finding is in line with recent studies (Ab Rahman, 2017; Muhamad Hanafi 2016; Bonk & Ockey, 2003; Caban, 2003; Kondo-Brown, 2002; Weigle, 1998; Lunz & Stahl, 1990; Wigglesworth, 1993; Lumley & McNamara, 1995). In their studies, the differences in the severity of the raters were significant but the infit MNSQ was within the permissible range. This indicated that the raters had multiple severities but the consistency of the raters was high. Although the raters had the same educational background and have more than 10 years of teaching experience, the raters were people who had their own characteristics, whether or not to be severe on a specific item or severe to a particular student. Effect of severity may exist because the rater consisted of various backgrounds. There was a rater whose character was severe and there was a rater whose character was lenient (Kondo, 2002). Raters can be lenient in some situations and severe in other circumstances (Prieto & Nieto, 2014). In addition, the raters also translated the scale of the rating with different meanings. According to Eckas (2011), the raters should have the same severity so that the assessment made on the student is fair.

The rater bias interaction findings in this study are consistent with past studies on the tendency of raters to be severe or lenient on the students and on the items in performance assessments. The findings from the recent study showed that there was a significant bias interaction between the raters on the items (Ab Rahman, 2017; Muhamad Hanafi 2016; Farrokhi et al., 2012; Schaefer, 2008; Eckes, 2005; Kondo-Brown, 2002; Lynch & McNamara, 1998; Wigglesworth, 1993) and bias interaction between the raters on the students (Ab Rahman, 2017; Farrokhi et al., 2012; Schaefer, 2008; Eckes, 2005; Kondo-Brown, 2002; Lynch & McNamara, 1998). According to Eckas (2011), the raters should have the same severity so that the assessment made on the student is fair. However, according to Lunz and Stahl (1990), the raters should not be trained to achieve the same severity. Many studies suggested that training should be provided to the raters to train them to achieve the same agreement in scoring (Ab Rahman 2017; Weigle, 1998; Barrett, 2001) and standardise their understanding of the scoring scales (Zhu et al., 2009). However, training cannot eliminate rater variability but can train raters to be more self-consistent (Lumley & McNamara, 1995). According to Engelhard (1996), the index obtained from the training was essential to filter out explicit rater scores, aiming to provide feedback to the raters, to monitor the continuous quality of the raters from time to time, and to evaluate the effect of training on raters. In addition to intensive training, Farrokhi et al. (2012) and Kondo-Brown (2002) recommended qualitative assessment to be carried out so as to understand why such a thing can happen.

5. Conclusion

The use of the proper rubrics helped the raters in giving consistent score to the students in producing a component installation of the electronic circuit project. Although the rubric is said to be a good instrument for measuring student competence, the rater effect still existed. Based on the results obtained, it can be concluded that the students had a high level of performance compared to rater severity and item difficulty. This indicated that the students mastering the exercises performed, but on the other hand showed that the rater was too lenient. There was severity effect at the individual level but did not exist at the group level. The raters were also biased towards items and students. Although the raters did not exercise the same level of severity when evaluating students, they had self-consistency in making the assessment. Even with the raters who had experience in teaching electronics, they also translated the scale of the rating with different meanings. Many studies suggested that training should be given to the raters to train them to reach the same agreement in scoring. However, raters cannot be trained to achieve the same severity but can be trained to be more self-consistent. Overall, the rubric cannot eliminate the effect of the raters directly but can give consistency to the raters so that they have the same understanding when translating the rubric. Performance assessment is very subjective and spreads to measurement errors if not controlled. There were various factors that influenced the effect of raters, among them were integrity factors, internal factors such as severity, halo, restriction of range, and central tendency. Raters are people who are exposed to emotions that can be varied.

Hence, producing a valid and reliable performance assessment is a difficult matter. However, using a rubric can at least reduce the measurement error gap contributed in the subjective assessment.

Acknowledgement
This research has been funded by Universiti Tun Hussein Onn Malaysia (UTHM) under PPG Grant Vot V007.

References

Abdullah, E. & Shukor, A. A. (2012). *Pentaksiran Prestasi dan Pentaksiran Rujukan Standard dalam Bilik Darjah*. Universiti Pendidikan Sultan Idris Tanjong Malim.

Ab. Rahman, A., (2017). Membina dan menentusah rubrik pentaksiran kompetensi amali kursus elektronik di Kolej Vokasional. Tesis Doktor Falsafah. Universiti Kebangsaan Malaysia.

Arter, J. & McTighe, J. (2001). S*coring rubric in the classroom: using performance criteria for assessing and improving student performance*. Corwin Press, INC.

Bonk, W. & Ockey, G. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, *20*(2), 238–252.

Caban, H. L. (2003). Raters group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, *21*(2), 1–44.

DeCotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance*, *19*(2), 247–266.

Eckes, T. (2005). Examining raters effects in TestDaF writing and speaking performance assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly: An International Journal*, *2*(3), 197–221.

Eckes, T. (2008). Raters types in writing performance assessments: A classification approach to raters variability. *Language Testing*.

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement : Analyzing and evaluating raters-mediated assessments*. Peter Lang.

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

Engelhard Jr., G. (1996). Clarification to "Examining raters errors in the assessment of written composition with a many-faceted Rasch model." *Journal of Educational Measurement*, *33*(2), 115–116.

Engelhard, G. (1996). Evaluating raters accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56–70.

Engelhard, G. (2013). *Invariant Measurement : Using Rasch Models in the Social Behavioral and Health Sciences*. Routledge Taylor & Francis Group.

Farrokhi, F., Esfandiari, R. & Schaefer, E. (2012). A Many-Facet Rasch Measurement of differential raters severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79–102.

Gipps, C. & Stobart, G. (2003). Alternative assessment. Student assessment and Testing (Vol3). 172-198

Iramaneerat, C. & Yudkowsky, R. (2006). How good are our Raters? Rater error in clinical skills assessment. *Education and Public Good: Interdisciplinary trend in Graduate Scholarship*.

Knoch, U., Read, J. & Von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*. *12*(1), 26–43.

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

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*. *21*(1), 1–27.

Linacre, J. (2002). What do infit and outfit, mean square and standardized mean? *Rasch Measurement Transaction*.

Lumley, T. & McNamara, T. (1995) Rater characteristics and rater bias: Implications for training. *Language Testing*,.

Lunz, M. & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, (13), 425–444.

Matsuno, S. (2009). Self, peer, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*. *26*(1), 075–100.

McNamara, T. F. & Adams, R. J. (1991). Exploring rater behaviour with rasch technique.

McQueen, J. & Congdon, P. J. (1997). *Raters Severity in Large Scale Assessment: Is it invariant?*

Montgomery, K., (2002) Authentic Tasks and Rubrics: Going beyond traditional assessments in college teaching. 50:1, 34-40, DOI: 10.1080/87567550209595870

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring raters effects using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement*, *5*(2), 189–227.

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

Muhamad Hanafi, N. (2016). Pembangunan dan validasi rubrik pentaksiran prestasi bagi mentaksir projek rekabentuk senibina di politeknik Malaysia. Tesis Doktor Falsafah, Universiti Pendidikan Sultan Idris.

Prieto, G. & Nieto, E. (2014). Analysis of raters severity on written expression exam using Many Faceted Rasch Measurement. 385–397.

Reddy, Y. M. & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448. doi:10.1080/02602930902862859

Saal, F. E., Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychomotor quality of rating data. *Psychological Bulletin*, *88*(2), 413–428.

Saal, F. E. & Landy, F. J. (1977). The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance*, *18*(1), 19–35.

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

Stokking, K., Schaaf, M. V., Jaspers, J. & Erkens, G. (2004). Assessment of student research skills. *British Educational Research Journal*, *30*(1), 93–116.

Till, H., Myford, C. & Dowell, J. (2013). Improving students selection using multiple mini-interviews with Multifaceted Rasch Modeling. *Academic medicine : Journal of the Association of American Medical Colleges*, *88*(2), 216–223.

Waugh, K. & Gronlund, N. E. (2013). *Assessment of students achievement*.

Weigle, S. (1998). Using FACETS to model raters training effects. *Language Testing*.

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

Wright, B. D., & Masters, G.N., (1982). *Rating Scale Analysis : Rasch Measurement.* Mesa Press. Chicago.

Zaghloul, A.R. M. (2001). Assessment of lab work : A three-domain model ; cognitive , affective , and psychomotor. *Proceedings of 2001 American Society for Engineering Education Annual Conference &Exposition*.

Zhu, W., Ennis, C. D. & Chen, A. (2009). Many-Faceted Rasch Modeling Expert Judgment in Test Development. *Measurement in Physical Education and Exercise Science*, (August 2013), 37–41.