Investigating Central Tendency in Competency Assessment of Design Electronic Circuit: Analysis Using Many Facet Rasch Measurement (MFRM)

Azmanirah Ab Rahman, Jamil Ahmad, Ruhizan Mohammad Yasin, and Nurfirdawati Muhamad Hanafi

***Abstract*—Rater plays an important role in awarding fair judgment to students. However, the difficulty to consider fairness to the student applies, especially for the assessment of competency in design electronic circuit. Therefore, the use of an analytic scoring rubric as a guide can reduce the error due to the nature of rubrics. This present research employs Many Facet Rasch Measurement (MFRM) to explore rater error focusing on central tendency effect. Participants comprised of a sample of nine experienced teachers who were employed to assess 68 students in their competency of Electronic Circuit Design process in Vocational College in Malaysia. Students were observed using four-point analytic rating scale. The data were collected and analyzed using FACET, a MFRM computer software program. The results were presented in two ways: at the group level and at the individual level. At the group level, information from the scale category statistics indicated central tendency effect; however, none of the separation statistics indicated such an effect. At the individual level, there are two raters that exhibit centrality**

***Index Terms*—Central tendency, many facet Rasch measurement (MFRM)**

I. Introduction

Practical work is one of the examples of performance based assessment that relies heavily on human judgment. However, this kind of assessment raises a variety of problems, especially in terms of validity and reliability since human judgment is subjective and uses a scale that is not dichotomized [1], [2]. In order to reduce human measurement errors and increase the validity and reliability to the decisions made in measuring students' abilities during practical work, a standard measurement process that includes a set of specific criteria and procedures that is valid and fair to all students should be conducted [3], [4]. An evaluation method that is claimed to be effective for fair judgment to students is by developing a standard rubric.

Rubric refers to a scoring guide used to evaluate the quality of students' constructed response. A rubric has four essential features: task description, scale, dimensions and descriptions of the dimensions. Task description is almost originally framed by the instructor and involves a performance of some sort by the student. The task can also apply to behavior, participation, use of proper lab protocols and behavioral expectations in the classroom. The scale explains how well or how poorly any given task has been performed. Terms used to describe the level of performance should be tactful and clear. The dimensions of the rubric present parts of the task simply and completely. Description of the dimensions means the quality definition represent of the quality of the performance for each rating scale [5].

Rating scale is a measurement instrument used by the rater to assigns ratees to positions along the continuum, denoting their relative ordering with respect to the trait being measured [6]. While rating scale provides information on the students' performance, they are unfortunately subject to various sources of bias and error. Irrelevant factors can influence human judgment process, such as individual raters and certain tendencies that may be exhibited [6]. The four most common rater errors are severity/leniency, halo effect, central tendency effect and restriction of range [7]. This research is focused on Central Tendency Effect.

II. Central Tendency Effect

There are several definitions of central tendency noted by many researchers. Central tendency is defined as the tendency to interpret and apply the scoring scale for category idiosyncratic. Raters tend to use the mid-scale category without properly assessing student performance [8]. DeCottis defined central tendency as a rater's unwillingness to go out on the proverbial limb in either direction, characterized by central ratings with little variability [9]. It is a special case of restriction of range [6], [7].

In the context of the MFRM, a rater is said to exhibit central tendency effect when he or she overuses the middle category, or middle categories of a rating scale while assigning fewer score at both the high and low ends of the scale [10]. The result is less variation in performance among students [11], [10]. According to [11], there are several factors that contribute to measurement error such as, 1) the rater failed to discriminate against students who were in a range between the lowest and highest. Therefore, the rater tends to leave marks on the mid-range 2) the examiner failed to distinguish between students' performance along the continuum, which does not understand the difference between each scale category. Thus, the examiner will allocate the same score to each student; 3) The rater did not have a strong background on the field or no training on the use of the rating scale during evaluation; 4) The examiner wants to be in a "play it safe strategy" for fear of being too soft or too lenient in awarding

Manuscript received November 20, 2015; revised June 2, 2016.

Azmanirah Ab Rahman and Nurfirdawati Muhamad Hanafi are with Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja Batu Pahat Johor, Malaysia (e-mail: azmanira@uthmedu.my, nurfirda@uthmedu.my).

Jamil Ahmad and Ruhizan Mohammad Yasin are with Universiti Kebangsaan Malaysia, 43650 Bangi Selangor Darul Ehsan, Malaysia (e-mail: jamil3191@yahoo.com, ruhizan@ukm.my).

marks. However [11] argued that this was not only due to the raters if many raters are likely to use the middle of the scale, there may also be a problem with the rating scale used in the instrument.

Centrality effect is a well-established phenomenon documented in various contexts. Many researches have been done on central tendency effect [12]-[15]. Leckie & Baird (2011) conducted a research to detect rater error in scoring essay writing in English for students of 14 years in England. The central tendency effect was present at an average level. Raters tend to over-score the poor essays and underscore a good quality essay. This occurs due to the diversity of experienced examiners. In their study, Farrokhi, Esfandiari and Dalili (2011) detected the central tendency effect of three types of rater, namely self rater, peer rater and teacher rater for English as second language. The results did not show any centrality effect for all three types of rater at either group or individual level. This is due to several strategies proposed by MyFord & Wolfe (2003) in their study. The rater was asked to give marks for "forced distribution", having them place a pre-specified number of points in each rating category and avoid monitoring during the test. Monitoring will impact the inspectors to use "play it safe strategy". Research by Knoch, Read & von Randow (2007) compared the effectiveness of online and face-to-face training assessing for the large-scale academic writing examination for students entering the University through the medium of English. Face-to-face raters tend to use the mid-range scale even after training. This is because they were monitored during the assessment process. Therefore, to reduce the impact of central tendency, it is better to avoid monitoring during the assessment process. The study by Kozaki (2004) found that 78% usage of the scale is located in the middle of the scale (scale 2 and 3). Two out of four raters are likely to contribute to the central tendency leniency to students who are less able and severe to students with high ability. This indicates that raters give the same marks to the students who have different abilities. Meanwhile, a study by Wolfe (2004) found that only 4% of raters tend to use middle scale category. The selection of experienced rater has been able to reduce the impact of central tendency.

III. Many Facet Rasch Measurement

Rasch models consist of an ideal type model for dichotomous, polytomous, partial credit, rating scale and many facet data. In this research, the focus is in Many Facet data. Many facet is designed to generalize the rating scale for rater-mediated assessment. Rater-mediated assessment includes a variety of assessments that require a rater or assessor to make judgments to assign score to a ratee [8]. Many facet is an extension of Rasch mesurement model, which suggests three facets: rater, ratee and traits, which interact to produce an observation [16]. Other than that, MFRM can also be used to identify rater errors in measurement with using rating scale model.

The effect of central tendency is presented in two ways: group level and individual level [10], [15]. At the group level, it can be detected using category scale, examinee separation statistics and criterion fit statistics. Category scale can be analyzed directly using Threshold Rasch Andrich. If frequency count and percentage of rating that raters assigned in middle scale category are higher than high end and low end, it indicates central tendency. Thus, there was an imbalance in the spread of ratings across the scale category [10]. More information about group level central tendency, according to [11], can be investigated by examinee separation statistics, which are fixed chi square, separation ratio, index separation ratio and reliability of separation index. 1) A fixed chi square test of the hypothesis shows that all ratees exhibit the same calibrated level of performance. A non-significant chi square value suggests a group central tendency effect. 2) Ratee separation ratio is to measure the spread of the ratee performance measures relative to the precision of those measures. A low separation ratio suggest a group central tendency effect; 3) Ratee separation index connotes the number of statistically distinct levels of ratee performance. A low ratee separation index suggests a group level central tendency effect; 4) Reliability of ratee separation index is a measure of the spread of the ratee performance measures relative to their precision. This index shows to what extent the raters are able to reliably distinguish among students in terms of their performance. A low reliability separation index suggests a group level of central tendency effect [11]. Criterion fit statistics provide another information to detect central tendency as well. Fit MNSQ index that is significantly overfit presents central tendency effect.

Criterion fit statistics is another source of information that may be used at the group level to identify central tendency effect. Infit and Outfit MNSQ indices that are overfit would contribute to the presence of central tendency effect [10]. As stated by [16], fit MNSQ value is between 0.5(lower level) and 1.5(upper level) is accepted. Fit value less than 0.5 is considered as overfit.

IV. Methodology

This research is conducted to investigate the group central tendency effect towards students' performance using rubric of competency in Vocational College. This study was conducted at three vocational colleges that offer the subject of Electronics Technology: Application of Electronic Circuits module (ETN 201). Three teachers from each college were selected to be raters. This is to make it easier for teachers to evaluate students. The sample consisted of 68 students. Raters were briefly introduced to the criteria of the rubric before they started implementing it in workshop. Rubric consists of 14 task descriptions using 4 rating scales. Each rating scale represents different criteria. For example, rating scale 1 represents poor, scale rating 2 represents moderate, scale rating 3 represents good and scale rating 4 represents excellent. The data were collected and analyzed using FACET, a MFRM computer software program.

V. Results

There are two ways to analyze data to detect central tendency. The first one is at group level and the second one is at individual level. Examining central tendency at the group level is easier than individual level because the researcher can directly observe from the Andrich Threshold table from rating

scale model.

At the group level, central tendency effect can be detected using scale categories (Table Threshold Andrich) and the statistical separation. Threshold Andrich scale category refers to the number of frequency count and percentage of use on mid-range ranking scale. Table 1 shows the frequency of use scale category. The lowest percentage of usage category (scale 1) and the highest category (scale 4) together was 31% whereas the percentage of usage of mid-scale (scale of 2 and scale 3) was 68%. Based on the findings, there is an irregular dispersion across each scale category, reflecting the impact of central tendency at the group stage.

TABLE I: SUMMARY OF CATEGORY STATISCIC

Score	Category Scale (%)	Outfit MNSQ	Threshold Measure
1	3	0.8	None
2	21	1.1	-2.30
3	47	1.0	-0.1
4	28	1.0	2.4

Central tendency effect can be identified using the statistical separation technique as in Table 2. The four indicators to track the central tendency are as follows: More information about central tendency is analyzed as below:

1) A fixed chi square test of the hypothesis that all ratees exhibit the same calibrated level of performance (that all ratees share the same performance measure, after accounting for measurement error). A non-significant chi square value suggests a group level central tendency effect [11]. The chi square value for is 758.9 with 67 degree of freedom is statistically significant ($p< 0.05$), suggesting that there is no group level central tendency effect.
2) Ratee separation ratio is to measure the spread of the ratee performance measures relative to the precision of those measures. A low ratee separation ratio suggests a group level central tendency effect [11]. From the result in Table 1, the separation ratio for ratee shows a high ratee separation ratio indicating three times larger than the precision. This indicator does not suggest a group level central tendency effect.
3) Ratee separation index connotes the number of statistically distinct levels of ratee performance. A low ratee separation index suggests a group level central tendency effect [11]. The ratee separation shows more than four statistically distinct levels of student performance. There is no evidence of central tendency because of high index value.
4) Reliability of ratee separation index is a measure of the spread of the ratee performance measures relative to their precision. This index shows to what extent the raters are able to reliably distinguish among students in terms of their performance. A low reliability of separation index suggests a group level of central tendency effect [11]. The reliability of ratee separation index shows the high degree of ratee separation reliability (0.92). This high reliability indicates that the rater can differentiate among ratees reliably.

Central tendency effects at the individual level are more informative and useful than at the group level. At the individual level, effects of central tendency can be detected using the statistical categories for each rater: MNSQ outfit and Threshold category of generating a hybrid model # 1. This information can be obtained from Table 3 and Table 4. Outfit MNSQ greater than 1.5 indicates the presence of effects of central tendency. Outfit MNSQ for rater 1 and rater 3 for middle scale category is larger than the expected value, indicating that there was more impact caused by the inspectors of central tendency than the other inspectors. To know in detail which rater is more likely to contribute central tendency effect, the average threshold is analyzed. The average distance threshold for rater 1 is 2.23, which is ([1.72 + 2.75] / 2) and rater 3 is 2.98, which is ([2.9 + 3:06] / 2). The large range reflects that the rater has more tendency to use the scale. This shows that rater 3 tends to contribute central tendency.

TABLE II: SUMMARY OF SEPARATION STATISTICS

Ratee Separation Statistic	
Homogeneity Index	758.9
Separation ratio	3.36
Separation (strata) index	4.81
Separation reliability	0.92

TABLE III: CATEGORY STATISTIC FOR RATER 1

Score	Category Scale (%)	Outfit MNSQ	Threshold Measure
1	5	0.9	None
2	19	1.5	-1.72
3	54	1.5	-0.63
4	23	1.1	2.35

TABLE IV: CATEGORY STATISTIC FOR RATER 3

Score	Category Scale (%)	Outfit MNSQ	Threshold Measure
1	2	0.6	None
2	28	1.6	-2.9
3	56	1.2	-0.16
4	14	1.2	3.06

VI. CONCLUSION

Based on the findings, central tendency effects can be traced at the group level and individual level. At the group level, information from the scale category statistics indicated central tendency effect; however, none of the separation statistics indicated such an effect. At the individual level, centrality effects are contributed by the two raters. This became possible because the first-time raters use a rubric to determine student performance. Examining using a rubric for the first time should be given adequate training in order to be able to distinguish the scale well. It is still difficult for the rater to discern any scale makes them less likely to leave marks on the middle scale. Scoring on a mid-scale is fairer to students in their response. Although the examiner is not monitored during the process of assessing such findings [14] during the online inspection but the raters felt safe to leave marks on the middle scale. Thus, the overall use of rubrics is to give marks to students who just give a good impression in reducing errors in measurement.

ACKNOWLEDGMENT

This paper is under the sponsorship of Universiti Tun

Hussein Onn Malaysia.

REFERENCES

[1] C. Gipps and G. Stobart, "Alternative assessment," 2003.
[2] T. J. Crooks, T. Kane, and A. S. Cohen, "T to the valid use of assessment," *Assess. Educ.*, vol. 3, no. 3, pp. 265-285, 1996.
[3] K. Stokking, M. V. Schaaf, J. Jaspers, and G. Erkens, "Assessment of student research skills," *Br. Educ. Res. J.*, vol. 30, no. 1, pp. 93-116, 2004.
[4] S. Messick, "Validity of psychological assessment. Validation of inferences from person's response and performance," 1995.
[5] D. D. Steven and A. J. Levi, *Introduction to Rubrics*, 2nd Ed. Stylus Publishing, 2013.
[6] C. M. Myford and E. W. Wolfe, "Detecting and measuring rater effects using many-facet Rasch measurement: Part I," *J. Appl. Meas.*, vol. 5, no. 2, pp. 189-227, 2003.
[7] F. E. Saal, R. G. Downey, and M. A. Lahey, "Rating the ratings: Assessing the psychomotor quality of rating data," *Psychol. Bull.*, vol. 88, no. 2, pp. 413-428, 1980.
[8] G. Engelhard, *Invariant Measurement: Using Rasch Models in the Social Behavioral and Health Sciences*, 2013.
[9] T. A. DeCotiis, "An analysis of the external validity and applied relevance of three rating formats," *Organ. Behav. Hum. Perform.*, vol. 19, no. 2, pp. 247-266, 1977.
[10] T. Eckes, *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*, Peter Lang, 2011.
[11] C. M. Myford and E. W. Wolfe, "Detecting and measuring rater effects using many-facet Rasch measurement: Part II," *J. Appl. Meas.*, vol. 5, no. 2, pp. 189-227, 2004.
[12] G. Leckie and J. A. Baird, "Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience," *J. Educ. Meas.*, vol. 48, no. 4, pp. 399-418, 2011.
[13] F. Farrokhi, R. Esfandiari, and M. V. Dalili, "Applying the many-facet Rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment," *World Appl. Sci. J.*, vol. 15, pp. 70-77, 2011.
[14] U. Knoch, J. Read, and J. V. Randow, "Re-training writing raters online: How does it compare with face-to-face training?" *Assess. Writ.*, vol. 12, no. 1, pp. 26-43, 2007.
[15] E. W. Wolfe, "Identifying rater effects using latent trait models," *Psychol. Sci.*, vol. 46, no. 1, pp. 35-51, 2004.
[16] J. Linacre, "What do infit and outfit MNSQ standardized mean," *Rasch Measurement Transaction*, 2002.

Azmanirah bt Abdul Rahman is a senior lecturer at Universiti Tun Hussein Onn Malaysia (UTHM). She has been teaching for 14 years in electronic engineering and technical and vocational education. She received a degree in electrical engineering (microelectronics), from Universiti Teknologi Malaysia (UTM) in 1999 and the master of technical education from UTHM in 2002. She is now pursuing PhD in Universiti Kebangsaan Malaysia (UKM) in the field of assessment and evaluation. Her research interest is assessment in engineering education and technical and vocational education.

Jamil bin Ahmad is currently a senior lecturer at the Universiti Kebangsaan Malaysia (UKM). He graduated from the same university in all three degrees; a bachelor degree in science in 1981, the masters in education in 1993 and the PhD in measurement and evaluation. His research interest lies his doctoral research in education specifically in the area of measurement and evaluation, course/program assessment and action research.

Ruhizan binti Mohamad Yasin is a professor at the Universiti Kebangsaan Malaysia and has been a lecturer since 1992. She graduated from University of the Pacific California USA with a bachelor degree in science (physics) in 1986. She later received her masters of science in teaching (physics) in Portland University, Oregon USA in 1988. She then decided to shift to the education discipline and received her diploma in education in UKM in 1990. After 8 years, she received her PhD in TVET from the University of Minnesota, USA. Her area of research interest is TVET, education for sustainable development and lifelong learning.

Nurfirdawati binti Muhamad Hanafi is currently a lecturer at Universiti Tun Hussein Onn Malaysia (UTHM) since 2004. She has been teaching almost 12 years in technical and vocational education. She received her degree in building surveying from Universiti Teknologi Mara, Malaysia in 2001 and the master of technical education from UTHM in 2002. Her PhD is in measurement education in Universiti Pendidikan Sultan Idris Shah (UPSI). Her research interest is in measurement and evaluation, instrument development and technical and vocational education.