How Assessment in Final Year Project Can Threatening the Validity?

Nur Afifah M.S, Nurfirdawati, M.H Azmanirah A.R, Nurul Hidayah L.A, Siti Wan Aisha W.A

Abstract*: An undergraduate project evaluation is measured using a rubric assessment. The tendency for the examiner to be biased is high. Otherwise, these subjective evaluations will threaten the validity and reliability of the assessment. Therefore, this study was conducted to investigate the effect of halo in the evaluation of undergraduate projects using analysis software program Multi Facet Rasch Measurement (MFRM) version 3.82.2. The methodology of this study is by using a design approach quantitative form. A total number of 98 students and 43 lecturers who serve as examiners of the Faculty of Electrical and Electronic Engineering of Tun Hussein Onn Malaysia University were involved in this study using purposive sampling. Halo effects were analyzed based on the objective of the study and detected using two methods, as a group and individually. In the analysis group there were four indicators examined (i) A Fixed Chi-Squared Test (Fixed) (ii) The trait separation ratio (iii) The trait separation index and (iv) The Reliability of the trait separation index. Meanwhile, in the analysis of individual level was examined by identifying significant bias interactions. From the study, it was found that studies show that there are halo effects at the individual level while halo effects cannot be detected at the group level. The findings of the study are to create that measurement is an important element in interpreting student performance. Although halo effects can be identified individually among examiners. But it does not affect the assessment because it does not exist in the group. Training and coordination need to be done to further develop the same level of understanding to improve student's ability to be translated to actual ability. Quality assessment will produce excellent and valuable human capital in the future.

Keywords: Validity, Reliability, Halo effect.

I. INTRODUCTION

The quality of University graduates is measured by their academic achievement. Holistic academic achievement is measured in a variety of ways. One of the most frequently used forms of lecturers is project-based assessment and examination. Examination through examinations is easier to do because it is more objective but different from assessment through more subjective projects. At the University in particular, to fulfil one of the Bachelor's Degree requirements, all students are required to undertake the Final Year Project (FYP) as one of the components of the study. The Bachelor Project is project-based learning and assessed based on performance assessment. Final Year Project employ the judgment of human raters using analytic scoring rubrics, and one problem that has been noted is that rater judgments involve a level of subjectivity that may hamper fair assessment (Hamp-Lyons & Kroll, 1996; Lumley, 2005; Van Moere, 2014). Students of similar ability may receive higher or lower scores based on the severity or lenience of their raters rather than on their ability level.

Performance assessment is an assessment that measures students' abilities and competencies in real-time situations (Baker & Linn 2002; Crooks, Kane & Cohen 1996; Gipps & Stobart 2003; Waugh & Gronlund 2013). However, performance appraisals are exposed to threatening validity and reliability (Gipps & Stobart 2003). According to Arter and Mctighe (2001), the answers given by students cannot be evaluated using machines because the students' answers are different and subjective. Poorly planned and practiced performance assessments cause frustration among students (Zaghloul 2001). Performance assessment not only involves an academic perspective, but also should be comprehensive, taking into account student skills. Final year projects are difficult to assess because they cover several aspects that need to be taken into account. Some of the aspects that need to be taken into account are the credentials given by the supervisor, the evaluation of the research report and the presentation (Ahmad et al., 2011).

Performance appraisals are difficult to measure because they are subjective and can invite examiner bias in assessments. There are many studies in western countries that have examined the impact of assessments on student performance such as the effect of firmness / softness effect, halo effect and central tendency (Eckes 2005; Farrokhi, Esfandiari & Schaefer 2012, Schaefer 2008). Examiners also tend to score different scores on the same student at different times, or two different assessors tend to score differently on the same student (Nitko 2001; Coffman 1971). In addition to assessors, rating scale also influences student scoring. The tendency of examiners to use mid-scale as a 'play it safe' strategy has led to a lack of diversity among students (Farrokhi, Esfandiari & Dalili 2011; Knoch, Read & Von Randow 2007; Kozaki 2004; Myford, & Wolfe 2004). In order to ensure that the justice of the lecturer needs to conduct a careful and careful assessment, evaluate with a careful heart (Ahmad & Awang, 2008).

Final year project are also performance assessments are assessments that are difficult to implement and require in-depth research by examiners.

Revised Manuscript Received on January 15, 2020

***** Correspondence Author

Nur Afifah Binti Mohd Sofian, Master student, Faculty of Education, Universiti Tun Hussein Onn Malaysia, Malaysia.

Nurfirdawati Binti Muhamad Hanafi,Senior lecturer Faculty of Education, Universiti Tun Hussein Onn Malaysia, Malaysia. Azmanirah

Binti Ab Rahman Senior lecturer Faculty of Education, Universiti Tun Husssein Onn Malaysia, Malaysia.

Nurul Hidayah Liew Abdullah, Senior lecturer Faculty of Education, Universiti Tun Husssein Onn Malaysia, Malaysia.

Siti Wan Aisha Binti Wan A'lauddin Master student, Faculty of Education, Universiti Tun Hussein Onn Malaysia, Malaysia.

Common in these assessments are the weaknesses of examiners assessing students' performance in terms of the effect of firmness, softness effect, halo effect, and central tendency. This examination error will have a serious impact on the credibility of Final Year Project evaluation system of an educational institution. According to Sapidin (2005), it increases stress and also engages in negative emotions among students and lecturers. According to Kohn, A. (2004) the education system in our Country is still critical of academic achievement in assessing an individual's educational level. This includes entry of students into grade 4, college and more. In addition, recruiting candidates to fill vacancies in the private and public sectors also reflects the importance of a student's academic achievement. Accordingly, the concern of lecturers in reducing bias when conducting assessments is very important in producing graduates who are qualified and therefore knowledgeable human capital.

RESEARCH OBJECTIVES

This study was conducted to see how the Final Year Project assessment threatened its validity. Validity is a test that measures something as well as possible (Anastati, 1988). The concept of validity refers to the whole thing to be measured and nothing else to measure (Thorndike 1991). Therefore, the objective of this study are to identify:

i) To identify the presence of halo effects at the group level using *Multi Facet Rasch Measurement*.

ii) To identify the presence of halo effects at the individual level using *Multi Facet Rasch Measurement.*

METHODOLOGY

The research design of this study takes the form of the quantitative research approach. This data will be collected using the assessment section as a research instrument and tested using the Many-Facet Rasch Measurement Model (MFRM). This research study was administered in Faculty of Electrical and Electronic Engineering of Tun Hussein Onn Malaysia. This research used purposive sampling. 98 students and 43 lecturers who serve as examiners had been purposely selected as the sample of study.

ANALYSIS OF DATA

In this study, data were collected using existing data from the Faculty of Electrical and Electronic Engineering of Tun Hussein Onn Malaysia University. The results of the 2017/2018 student Final Year Project were taken from the faculty's Final Year Project coordinator. After obtaining the data, the data were analyzed using Rasch Measurement Multi Facet software version 3.82.2 to identify the presence of halo effects at the group level and the presence of halo effect at the individual level. In the analysis group there were four indicators examined (i) A Fixed Chi- Squared Test (Fixed) (ii) The trait separation ratio (iii) The trait separation index and (iv) The Reliability of the trait separation index. Meanwhile, in the analysis of individual level was examined by identifying significant bias interactions.

FINDINGS AND DISCUSSION

The findings and discussion of the research was analyzed based on the research objective.

In the context of the MFRM analysis, the halo effect implies the tendency of the examiner to assign the same assessment rate to different feature sets. An examiner who has a halo effect in himself cannot make any difference to these characteristics and thus gives the same value across the traits. Researchers use any hybrid model to analyze the data. The results are made up of two stages, namely the group level and the individual level.

Linacre (personal communication January 20, 2003) stated that another method that can be used to detect halo effects is by using aspect. This idea is used by researchers to analyze behavior examiners. This analysis can show and suspect halo effects. In this case, the effect of the halo exists on the examiner when the researcher takes all the properties of the same difficulty (that is, usually 0). Raters showing rates are rater showing halo effect.

As a result, the trait reflects differences in the aspect of the difficulty. The obvious deficits in trait difficulty indicate that the examiner is unable to distinguish the traits and give the same rating. Using the rating scale model shows several halo levels that give priority to feature measurement

1) Group-Level Statistical Indicator

In the analysis using Multi Facet Rasch Measurement (MFRM), there are two methods for detecting halo effects that are group statistical analysis statistics and individual level analysis statistics. The group has four indicators that are examined

i) A Fixed Chi- Squared Test (Fixed)

Table 1 shows the results of the statistical analysis at the group level obtained from the analysis using the FACET program. The fixed chi-squared test was to test the hypothesis that all trait share the same level of difficulty. Based on table 1 the mean squared value is 224.1 with df = 12 indicating a significance value of $p < .05$. The results showed that there was no presence of halo effect in the measurements at the statistical analysis at the group level.

ii) The trait separation ratio

The trait separation ratio. This ration is a measure of the spread of the traits difficulty measure relative to the precision of those measure. A low trait separation ratio suggest halo in the rating. The trait separation ratio in Table 1 shows 4.30 indicating that the difficulty of the item is 4 times greater than the measurement accuracy. This indicator does not reflect the impact of halo in statistical analysis at the group level.

iii) The trait separation index

The trait separation index. This indicator indicates the level of difficulty of the item. This indicator connote the number of statistically distinct levels of trait difficulty among the traits included in the analysis.

A low trait separation index show halo in the ratings. Referring to table 1 shows that the trait separation index is 6.07 showing that there are over 6 different items of difficulty level statistically and directly this indicator means no halo effect in group level statistical analysis.

iv) The Reliability of the trait separation index

Reliability of the index of separation of properties. It gives an overview of the properties of the properties separated by their difficulty level, it also shows how well the examiner can distinguish the properties present. The low reliability coefficient index shows the halo effect. The confidence level of the separation of attributes (1.00) indicates that the mediator can reliably distinguish between the characteristics. The indicator automatically tells us that there is no halo effect on the group level in this data set. According to table 1 the reliability index of the examiner is 0.95. High levels of reliability suggest that it is not possible to detect halo at this stage.

Table 1: Trait Measurement Report from an Analysis Using the Rating Scale Model – Halo Stimulation

Model	Standard Deviation(S.D)	Separation	Strata	Reability
Model, Populn: RMSE	0.63	4.12	6.83	0.94
Model, Sample: RMSE	0.65	4.30	6.07	0.95
Model, Fixed (all same) chi-square: 224.1 d.f.: 12 significance (probability): .0.0				

2) Individual-Level Statistical Indicator

Individual-level statistical indicator. For each rater that included in an analysis using the rating scale model or any hybrid model, Facet provide measure of the consistency of the rater rating with the MFRM expected rating. In the individual-level analysis, the effect of halo was examined by identifying significant bias interactions. Based on the report of the bias interaction obtained through " Table 13.3.1 " in the analysis of the FACET program, it was found that there were 185 bias interactions between examiners and students in this section. Of the total, only 21 bias interactions identified were significant with p values <0.05 and t values greater than + 2. Figure 1 shows that the percentage of bias interactions between examiners and students was 11% significant. This finding shows that the presence of halo effects in individual-level statistical analysis is 11%. The findings obtained in this study are in line with previous studies (Farrokhi, & Esfandiari 2011); Kozaki 2004; Yorozuya, & Oller 1980). In the study of Farrokhi, & Esfandiari (2011) hello effects exist due to inexperienced examiners, inadequate training and no response given after training. However, according to the study of Knoch et al (2007) no significant change occurred despite training provided.

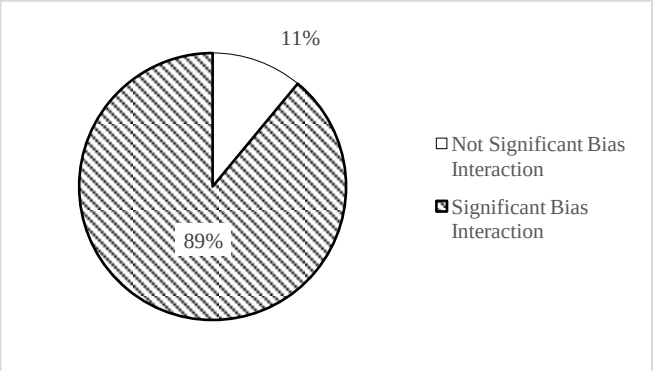

Figure 1: Bias Interaction Examiner- Student

In table 2 (Table 7 of *facet* output), it is show that rater 29 exhibit greater misfit when compared to the others rater. The infit and outfit mean-square indices for rater 29 (1.56 and 1.53, respectively as shown in column 7 and 8) are larger than those for remaining raters. It is also important to note that rater 29 does not exhibit an undue amount of randomness.

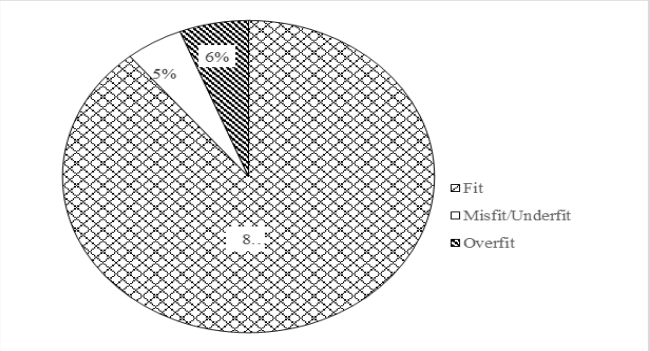

Figure 2: Fit value analysis

The next indicator examined by the researcher is to characterize the " fit " value for the biased interaction between the item and the examiner. Figure 2 shows the percentage of " fit 'values analyzed based on 19 significant bias interactions. From the figure it was found that 89% of the bias interactions met the Rasch Model (0.5 logit- 1.5 logit) while 6% the bias interaction was " misfit or " under fit " (> 1.5 logit) and the remaining 5% bias interaction was " over fit ". '(<0.50 logit).' 'Misfit' 'is more problematic than' 'over fit because' 'misfit does not alter the substantive meaning of a given measure and threatens the validity of the measurement system (Wright and Linacre, 1994; My ford & Wolfe, 2003). This finding means that only about 11% contributed to the scoring of this section. Although, only 11% of it still has a halo effect. Therefore, training should be given to the examiner to reduce the effect of the halo.A study by Iramaneerat, & Yudkowsky (2007) also detected halo effects using mfrm. The halo effect is detected by an inconsistent fit value. There was a halo effect in his study and he suggested that the examiner be given additional instructions to avoid the effect of the halo.

II. CONCLUSION

The halo effect can affect student test grades and it also has the potential to ruin student final test scores. It happens when humans are taken as assessors. (McNamara, 2000) judge-awarded ratings (Linacre, 2004). The halo effect is the tendency of the examiner to give the same score to the student even though the given trait has the same level of difficulty. According to Myford & Wolfe, 2004 examiners who exhibited halo effects were unable to identify good features by giving students similar scores. The halo effect as defined in the literature is the carry-over from one judgment to another that is, assigning similar ratings to rates across items Rubric was used to identify the halo effect present in this study.In addition, the halo effect is also a subset of stimulus features that contribute to the construction of variants that will cause test scores to be affected. This halo effect should be minimized, as students will be the main victims of the assessment. Examiners cannot conceptualize students' actual capabilities. In order to minimize this problem the examiner should be given training, experience and feedback. The contributing factor to the halo effect is due to the examiner's failure to make the difference criterion by making public lectures on Eckes' student achievement (2011). There are a few methods that we can do to reduce the bias when detected. One of the way is, provide rigorous training coupled with monitoring. The training will give more knowledge to the raters. Another helpful way is instruction, which might be dispense with the need for rater training (Saito, 2008). Instruction might provide raters with clear and explicit assessment criteria or might involve co-creation and negotiation of rating scales with raters. Other than that, feedback to raters has also proved to be helpful, especially if it is do in the long time (Knoch, 2011). Another reason for the presence of bias in the present study might be lack of feedback from the rater.The assessment is to look at the development of student knowledge. Good assessment tools are very important in a measurement because the accuracy of the measurement of the inference depends on the functionality of the measuring device. Therefore, educator should be aware that quality assessment is the only assessment that provides students with justice. The use of rubrics requires a clear understanding as each examiner has different understandings. Therefore, each examiner needs to have a clear knowledge of each item used so that the examiner has the same understanding. In order to obtain valid measurement results and assessments, the examiner must also be a credible examiner when making markings so that fair and equitable values can be provided to the student. Integrity should be emphasized in the teacher curriculum. If all the examiners have the integrity then the students' true abilities are based on the criteria they have created.

APPENDIX

Table 2: Rater Measurement Report from an Analysis Using the Rating Scale Model – Halo Stimulation

Obsvd average	Fair(M) Average	Measure	Model S.E	Infit MnSq	ZStd	Outfi MnSq	t ZStd	Estim. Discr	Correlation PtMea	Correlation PtExp	Nu	Rater
3.28	3.04	3.62	0.41	0.98	0	2.85	1.9	0.77	0.42	0.48	24	A024
4.08	4.05	3.51	1.07	0.71	-0.1	0.3	-0.5	1.22	0.71	0.23	25	A025
1.58	3.23	2.55	1.07	0.98	0.2	0.59	0	1.06	0.35	0.23	41	A041
3.5	3.41	2.23	0.27	0.82	-0.9	0.8	-0.6	1.23	0.61	0.53	17	A017
3.97	3.7	2.14	0.26	0.92	-0.4	0.95	-0.1	1.07	0.59	0.55	18	A018
3.49	3.51	1.77	0.29	0.94	-0.2	0.92	-0.3	1.07	0.43	0.38	9	A09
3.9	3.75	1.47	0.21	0.92	-0.5	0.92	-0.5	1.1	0.64	0.6	14	A014
3.69	3.88	1.13	0.31	0.61	-2.1	0.58	-2.2	1.4	0.78	0.54	2	A02
3.81	3.84	1.1	0.27	1.07	0.4	1.05	0.3	0.91	0.26	0.35	13	A013
3.37	3.24	1.05	0.27	0.78	-1.2	0.75	-1.2	1.22	0.67	0.53	26	A026
3.54	3.56	1.02	0.41	1.31	0.9	1.42	1	0.64	0.58	0.73	8	A08
2.44	3.4	0.9	0.39	0.77	-1.2	0.64	-1.3	1.39	0.62	0.36	32	A032
3.79	3.97	0.63	0.33	1.1	0.5	1.12	0.5	0.9	0.27	0.38	4	A04
3.44	2.97	0.58	0.37	0.81	-0.7	0.8	-0.8	1.31	0.8	0.76	7	A07
3.33	3.38	0.49	0.26	0.8	-0.9	0.78	-1	1.17	0.67	0.57	19	A019
3.26	3.3	0.48	0.2	0.73	-1.8	0.67	-2.1	1.34	0.82	0.74	1	A01
3.56	3.52	0.45	0.2	1.01	0	1	0	0.99	0.6	0.61	28	A028
3.52	3.57	0.07	0.18	0.97	-0.2	0.95	-0.3	1.04	0.47	0.44	40	A040
3.9	4.15	-0.03	0.3	1.07	0.3	1.07	0.3	0.91	0.33	0.41	27	A027
3.6	3.64	-0.07	0.29	0.96	-0.3	0.87	-0.4	1.11	0.52	0.5	34	A034
3.65	3.68	-0.11	0.24	1.12	0.6	1.12	0.6	0.87	0.36	0.47	16	A016
4.13	4.2	-0.2	0.3	0.91	-0.3	0.9	-0.4	1.13	0.54	0.46	23	A023
3.62	3.9	-0.25	0.31	0.73	-1.2	0.79	-0.8	1.23	0.87	0.83	5	A05
3.64	3.7	-0.36	0.38	0.98	0	1.1	0.4	1	0.45	0.46	20	A020
3.21	3.37	-0.38	0.17	1.56	2.6	1.53	2.5	0.62	0.56	0.69	29	A029
3.3	3.41	-0.49	0.2	1.16	0.9	1.25	1.4	0.8	0.33	0.48	30	A030
3.83	3.76	-0.65	0.39	1.22	0.9	1.92	1.8	0.61	-0.03	0.29	3	A03
4.35	4.18	-0.67	0.46	1.22	0.7	0.99	0.1	0.79	0.55	0.65	6	A06
3.37	3.18	-0.67	0.23	0.77	-1.7	0.66	-2.1	1.23	0.7	0.6	31	A031
3.12	3.28	-0.72	0.41	0.93	-0.1	0.92	-0.1	1.07	0.43	0.33	35	A035
3.52	3.43	-0.99	0.29	0.88	-0.6	0.87	-0.6	1.16	0.6	0.53	38	A038
3.71	3.98	-1.05	0.27	0.79	-1.1	0.8	-0.9	1.21	0.64	0.51	22	A022
3.42	3.47	-1.07	0.2	0.97	-0.1	0.97	-0.1	1.04	0.44	0.43	39	A039
2.51	4.26	-1.19	0.3	1.15	0.6	1.25	0.9	0.83	0.33	0.47	15	A015
3.24	3.26	-1.29	0.25	1.18	1.2	1.13	0.7	0.81	0.3	0.45	37	A037
3.36	3.38	-1.29	0.21	1.03	0.2	1	0	0.96	0.47	0.5	43	A043
2.99	2.99	-1.47	0.21	1.12	0.8	1.14	0.9	0.89	0.62	0.67	33	A033

3	3.51	-2.17	0.29	0.99	0	1.02	0.1	1	0.66	0.66	36	A037
3.87	3.94	-2.26	0.49	0.94	0	0.71	-0.5	1.09	0.32	0.19	11	A011
3.58	3.54	-2.34	0.22	1.25	1.7	1.27	1	0.73	0.52	0.63	42	A042
2.65	4.85	-2.54	0.49	0.87	-0.2	0.86	-0.3	1.23	0.51	0.36	10	A010
3.94	3.97	-2.95	0.61	0.98	0.1	0.66	-0.2	1.04	0.25	0.19	21	A021
3.45	3.63	0	0.34	0.98	-0.1	1	-0.1		0.51		Mean	(Count: 42)
0.49	0.38	1.52	0.19	0.18	1	0.4	1		0.18		S.D	(Population)
0.5	0.39	1.54	0.19	0.19	1	0.41	1.1		0.18		S.D	(Sample)
Model, Populn: RMSE .39 Adj (True) S.D. 1.47 Separation 3.78 Strata 5.37 Reliability .93												
Model, Sample: RMSE .39 Adj (True) S.D. 1.49 Separation 3.83 Strata 5.44 Reliability .94												
Model, Fixed (all same) chi-square: 839.9 d.f.: 41 significance (probability): .00												
Model, Random (normal) chi-square: 36.8 d.f.: 40 significance (probability): .61												

ACKNOWLEDGMENT

We would like to acknowledge the helpful advice supervisor, Dr Nurfidawati Binti Muhamad Hanafi, for his supervision and constant support. His invaluable help of constructive comments and suggestions throughout this journey works have contributed to the success of this paper. Not forgotten, my appreciation to my co-supervisor, Azmanirah Binti Ab Rahman for his support and knowledge regarding this topic.

I would like to express my appreciation to University Tun Hussein Onn Malaysia, who had done the process of publishing this paper under the PPG Research grant with Vot. No. K017.

REFERENCES

1. Ahmad, A. & Awang, M.I (2008). Pengukuran Dan Penilaian Pendidikan. Kuala Lumpur : Dewan Bahasa Dan Pustaka
2. Ahmad, R. R., Suradi, N. R. M., Majid, N., Shahabuddin, F. A., Rambely, A. S., Din, U. K. S., & Ali, Z. M. (2011). The Role of Final Year Project in The School of Mathematical Sciences in Human Capital Development. Procedia - Social and Behavioral Sciences, 18, 450–459. https://doi.org/10.1016/j.sbspro.2011.05.066
3. Arter, J. & McTighe, J. 2001. Scoring Rubric in the Classroom: Using Performance Criteria for Assessing ang Improving Student Performance.Corwin Press, INC
4. Anastasi, A. (1988). Psychological testing (6th ed.). New York, NY, England: Macmillan Publishing Co, Inc
5. Eckes, T. 2011. Intriduction to may facet rasch meaurement: Analyzing and Evaluating Rater-Mediated Asesement. Peter Lang.
6. Farrokhi, F., Esfandiari, R. & Schaefer, E. 2012. A Many-Facet Rasch Measurement of Differential Rater Severity/Leniency in Three Types of Assessment. JALT Journal, 34(1), 79–102
7. Gipps, C. & Stobart, G. 2003. Alternative Assessment. Iramaneerat, C. & Yudkowsky, R. 2007. Rater error in a clinical. Evaluation and the health Professions, 266-283.
8. Kohn, A (2004) Inilah Sekolah Idaman Pelaman, Siri Pembangunan Profesion Perguruan, Kuala Lumpur, ITBM
9. Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior—a longitudinal study. Language Testing, 28, 179-200.
10. Knoch, U., Read, J., & von Randow, T. (2007). Re-training writing raters online: How does it compare with face-to-factraining? Assessing Writing 12.2, 26–43.
11. 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.
12. Assessment: An overview. College ESL, 6(1), 52–72 assessment: An overview. College ESL, 6(1), 52–7Baker, F. B. (2001). The basics of item response theory. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation. Original work published in 1985. Retrieved from http://echo.edres.org:8080/irt/baker/Eckes, T. 2005. Examining raters effects in TestDaF writing and speaking performance assessments:A many-facet Rasch analysis. Language Assessment Quarterly: An Internation
13. Evaluation, methodology, and interdisciplinary themes (pp.
14. Evaluation, methodology, and interdisciplinary themes (pp.
15. Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writingHamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. College ESL, 6(1), 52–72Linn, R.L., (2001). Assessments And Accountability (Condensed Version). PracticalAssessment, Research & Evaluation, 7(11). Retrieved September 20, 2008 fromhttp://PAREonline.net/getvn.asp?v=7&n=11 . This paper has been viewed 44,339times since 116/2001
16. Linacre, J. M. (2004). Optimizing rating scale effectiveness. In., E. V., Smith, Jr, & R. M. Smith. (Eds.). Introduction to Rasch model. Maple Grove, Minnesota: JAM press, 258-278.
17. Lumley, T (2005). Assessing second language writing: The rater's perspective. Frankfurt am Main: Peter Lang.
18. McNamara, T. F. (2000). Language testing. Oxford, UK: Oxford University Press 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.
19. Myford, C. M. and E. W. Wolfe. (2004b). Detecting and measuring rater effects using many-facet Rasch measurement: Part II.
20. Nitko, A. J. (2001). Educational assessment of students (3rd ed.). Upper Saddle River, NJ: Merrill. Oller, 1980, Language Tests at School. London: Longman
21. Sapidin, A.M. (2005, Ogos 18). Stress Guru Membimbangkan: Akibat Banyak perubahan dan Tugas Yang Keterlaluan. National Union Of The Teaching Profession (NUTP) Utusan Malaysia , ms 12 Sebagai Teras Pembangunan Staf Berkualiti Jurnal Akademik UPM, 3(9),25-27
22. Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. Language Testing 25.4, 465–93
23. Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. Language Testing, 25, 553-581.
24. Thorndike, R. M., Cunningham, G. K., Thorndike, R. K., & Hagen, E. P. (1991). Measurement and evaluation in psychology and education (5th ed.). New York: Macmillan.
25. Van Moere, A. (2014). Raters and ratings. In A. J. Kunnan (Ed.), The companion to language assessment volume III: Evaluation, methodology, and interdisciplinary themes (pp. 1358–1374). Malden, MA: Wiley Blackwell
26. Waugh, C. K., & Gronlund, N. E. (2013). Assessment of Student Achievement (10th ed.). New York: Pearson
27. Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370-371
28. Yorozuya, R. & J. W. Oller, Jr. (1980). Oral proficiency scales: Construct validity and the halo effect. Language Learning.

AUTHORS PROFILE

Nur Afifah Binti Mohd Sofian is a Master's degree student in Technical and Vocational Education, Faculty of Education from University Tun Hussein Onn Malaysia, Malaysia (2018-2019). She graduated on her first degree ob B.(E.d) Hons in Vocational (Building Construction) from University Tun Hussein Onn Malaysia. Her current research study is in assessment and evaluation education. She focus on examiner bias in assessing undergraduate projects using analytical rubrics among University Tun Hussein Onn Malaysia students. Her previous research study during her first degree is also in assessment and evaluation area which is more concentrated on practice formative assessment in the environment of vocational lecturers around Johor state. Her experience in education world is about 6 years.

Nurfirdawati Binti Muhamad Hanafi,is a senior lecturer in the department of Engineering Education at the, Faculty of Education from University Tun Hussein Onn Malaysia and a leader Focus Group (FG)Testing, Evaluation, Assessment And Measurement (TEAM). She received her PhD from Sultan Idris Education University. Her research study is in a assessment and evaluation education studio based learning and architecture education. Her areas of research interest is on studio based learning, assessment and evaluation in education and architectural education. In addition, her innovation field of interest is in new approach in teaching and learning education and using assessment as a guidance in learning process. She also is author for more than 10 publication and book which is focusly in assessment and evaluation education.

Azmanirah Binti Ab Rahman is a senior lecturer in the department of Engineering Education at the, Faculty Of Education from University Tun Hussein Onn Malaysia. She received her PhD in assessment and evaluation from National University Of Malaysia in year 2017. She received her bachelor's degree in electrical enginering and M.(Ed) in vocational, in 1999 and 2002, respectively from University of Technology Malaysia.Her research focused on assessment and evaluation education. Her field of expertise is on Technical and Vocational Education. She received professional societies from Malaysia Board Of Technologist (MBOT) as a graduate technologist in 2018. She started her career as a lecturer since 2004 and have been teaching for more than 15 subject in electrical program. She has been an author for more than 15 publication and jurnal in variaties of area.

Nurul Hidayah Liew Abdullah is a senior lecturer in the department of Engineering Education at the, Faculty of Education from University Tun Hussein Onn Malaysia. She received her PhD in assessment and evaluation from University Tun Hussein Onn Malaysia in year 2015. Her field of expertise is in educational administration. Besides, her areas of research interest is on management of education according to the 21st century mold of environmental education and teaching and learning management in and out of the classroom. In addition her innovation field of Interest is in learning and teaching innovation.Her research focused on technology management. She start her career as a teacher at Sek. Men Ker Bahasa Malaysia Saratok Sri Aman Sarawak in year 1993. Her experience in education world is about 26 years. She has been an author for more than 10 publication and jurnal in variaties of area.

Siti Wan Aisha Binti Wan A'laudden is a Master's degree student in Technical and Vacational Education, Faculty of Education from University Tun Hussein Onn Malaysia, Malaysia (2018-2019). She graduated on her first degree ob B.(E.d) Hons in Vocational (Building Construction) from University Tun Hussein Onn Malaysia. In 2013, she receive her diploma in civil engineering field from polytechnic Kota Bharu. Her current research area of study is on assessment and evaluation education focusing currently on teachers candidate assessment in teaching practicum.

.