Investigating reliability and validity of student performance assessment in Higher Education using Rasch Model

Tengku Zatul Hidayah Tengku Petra¹, Mohd Juzaiddin Ab Aziz²

¹Center For Artificial Intelligence Technology, Faculty of Information Science and Technology, National University of Malaysia, 43600 Bangi, Selangor, Malaysia
²Center For Software Technology And Management, Faculty of Information Science and Technology, National University of Malaysia, 43600 Bangi, Selangor, Malaysia

*tg_zatul@yahoo.com

Abstract. The quality and outcomes of higher education institutions are gaining tremendous levels of interest due to the competitive environment nowadays. The process of teaching, learning, and evaluation of students indicates the students and programme performance in higher education. However, measurement of the students’ performance from the assessment remains vague when arithmetical methods based on raw scores were used. For the past several decades, a few studies have been conducted on the impact of assessment implemented towards students and programme performance. In this study, the authors investigated the performance of student assessment implemented by institutions using the result of student learning outcomes in terms of attributes’ reliability and validity. The student assessment result was divided into six attributes (knowledge, problem solving, communication, group work, lab work, and design skill) as the weighting factors for evaluating students’ performance. The test showed that the assessment implemented was valid and reliable, thus appropriate and effective in measuring students’ achievement and programme performance.

1. Introduction

Assessment is a systematic process that plays a significant role in effective teaching and learning. The process starts with the identification of learning outcomes and ends with a judgment to identify whether those outcomes have been attained [1]. Assessment is a general term that includes all procedures used to obtain information about student learning [1]. Evaluation occurs when a score is assigned after the completion of any assessment method or learning activity and focuses on ‘better or worse than’, ‘higher or lower than’, and ‘good or bad’ comparisons among students [2].

Assessing student performance requires a rigorous assessment in which the evidence of learning is collected in a systematic way. It encompasses two types of assessment, namely formative and summative, depending on the use of the results [3]. Formative assessment measures any progress in order to meet the goals and objectives, while summative assessment evaluates the work at the end of the learning process. These assessments can be measured directly or indirectly. Examples of direct assessment are final examinations, quizzes, assignments, and capstone projects, while indirect assessments include conducting student surveys.

Considering the crucial role of assessment in the teaching and learning process, it is imperative that all types of assessment fulfill the basic requirement for validity, reliability, and usability [1]. Validity refers to the adequacy and appropriateness of the interpretations made from assessment according to a particular use. For instance, if the results are to be used as a measure of students’ communication skill,
it is suggested for the interpretations to be based on evidence that the scores are solely reflected on communication skills and not affected by irrelevant factors. Reliability refers to the consistency of assessment results. If similar scores are obtained when the same assessment procedure is used on the same students at two different occasions, it can be concluded that the results have a high degree of reliability from one occasion to another. Similarly, if different lecturers mark students’ answer on the same assessment task and obtain quite similar results, it can also be concluded that the results have a high degree of reliability. Usability refers to the practicality of the process. It should be economical in terms of time and money, easily administered and scored, and produce results that can be interpreted and applied by learning institutions [1].

This research intended to analyse two items, namely the reliability and validity of the assessment used to measure student performance. In fact, validity and reliability have different meanings with different implications for student assessment. Validity refers to how well the assessment instrument measures what it is intended to measure. In contrast, reliability refers to consistent results over time. Reliability is necessary, but not sufficient to establish validity. For example, if a student weighs herself four times on a scale and obtains the values 64.0, 63.6, 63.8, and 64.0, then it can be said that the scale is reasonably reliable since the weights are consistent. Nevertheless, she actually weighs 70 kilogrammes and not 64; therefore, the scale measurement has little validity.

2. Methods of Assessment

Traditionally, institutions use actuarial data such as graduation rate and student-faculty ratio, ratings of institutional quality, student surveys, peer reviews, and direct measurement of student skills and knowledge to demonstrate efficacy [4-5].

Another method of measuring performance is collecting data on course grades (e.g. quizzes and examinations) and student work products (e.g. portfolios and concept maps) [4]. In the evaluation process of student work, problems arise when it comes to the question of who actually did the work and the validity and reliability of such approaches [4]. Course grades are the proper way to measure performance; however, they are specific to certain lecturers and non-standardised tests and are difficult to compare results even across programmes within a faculty.

The use of rubrics provides another way to address these issues [6]. Rubric is a scoring guide to assist the examiner when making a judgment about the quality of student work through a set of evaluative criteria [7]. General problems of rubric are low consistency among different raters and time-consuming.

Besides, Kumari et al. [8] used fuzzy logic reasoning to evaluate student performance by considering academic results as well as their personality traits. The attributes considered to measure academic performance were attendance, self-learning, examination, and lab result, while project skills, presentation skills, extracurricular activity/co-curricular activity (ECA/CCA), and attitude were used to measure personality development.

Assessment plays an important role in student performance measurement. An ideal measurement should be able to make use of the test score for some purposes (valid), accurate (reliable), have a stable frame of reference in terms of comparing different students, provide a linear measure (map the hard test scores onto the easy test scores), can give meanings to the score (relate to actual students’ ability), and detect misfits [9]. Therefore, the selection of proper tools for assessing student performance is very important.

This study aimed to explore a Rasch model for assessing validity and reliability of student assessment methods that have been implemented by various researchers and institutions using the result of students’ learning outcomes. The current research on Rasch model in student assessment involves the determination of a task whether it is easy or difficult and to classify students’ ability based on the score [10-14]. Others use the Rasch model to study the validity and reliability of the items [12-13,15-16].

2.1. Instrument
The Rasch model is a particular Item Response Theory (IRT) model. It indicates the success probability of a person on an item depending on the difficulty of the item and the ability of the person [9]. The Rasch model transforms raw scores, namely ordinal data, into a linear interval scale data [11]. Usually, the success probability on an item for people with different abilities is plotted as an “item characteristic curve” (ICC) as shown in Figure 1.

![Figure 1. Example of item characteristic curve.](image)

From Figure 1, it can be seen that high achievers have the probability of success close to one and low achievers close to zero. Meanwhile, average ability students have the probability of success at 0.5. The dotted line shows the probability of success on the item at each ability level. The Rasch model would scale the ability according to the success (or a correct response). The smaller portion of success, the higher the difficulty of an item. For the purpose of this study, the Winsteps software was employed to perform the Rasch analysis. From the analysis function in the software, the result of students and item reliability was obtained. This value could determine the reliability of the overall student assessment implemented through the result of students’ learning outcomes and detect any problems in the detailed results of individual analysis. The observed variance and error variance values were used in the calculation to find the reliability level, as shown in Equation 1.

\[
\text{Reliability Score} = \frac{\text{Observed Variance} - \text{Error Variance}}{\text{Observed Variance}} \quad (1)
\]

The common statistical measurement accepts the range of reliability score between 0.5 to 1.0 as reliable and acceptable. However, to be more valid and highly accepted, scores between 0.8 and 1.0 were used in this study. This is equivalent to the common measurement scheme in education that considers scores between 80%–100% as grade ‘A’.

2.2. Attributes and participants.

The result of students’ learning outcomes through assessment was divided into six attributes, which are: i) knowledge; ii) problem-solving skill; iii) communication skill; iv) group work; v) lab work; and vi) design skill. These six attributes were the weighting factors for evaluating student performance. Various types of assessment were used to measure student performance in these attributes, such as final examination, tutorial, report, and presentation. For the purpose of this study, 60 students from the Civil Engineering programme (Year 4) in National University of Malaysia (Universiti Kebangsaan Malaysia) were analysed. Table 1 shows the assessment used in this programme.
Table 1. Selection of input attributes.

| Input Attributes      | Assessment Methods                                      |
|-----------------------|---------------------------------------------------------|
| Knowledge             | Research proposal, tutorial, final report, thesis (literature review) |
| Problem Solving       | Final report, proposal, thesis (introduction and conclusion), tutorial, project, mid semester and final examination |
| Communication         | Report, proposal, presentation rubric and viva          |
| Group work            | Peer assessment                                         |
| Lab work              | Report, thesis                                          |
| Design & Experiment   | Proposal, thesis (method, result, discussion)           |

Table 2 exhibits an example of student performance report for knowledge attribute. Each attribute was given the percentage that indicated the students’ performance on those attributes. For the purpose of analysis using the Winsteps software, the percentage scores were converted into the rating scale of 1<60, 2≥60, 3≥70, 4≥80, 5≥90.

Table 2. Example of students’ performance report.

| Students | Research proposal mark | Final design report mark | Tutorial mark | Thesis evaluation (lit. review) mark | Overall total mark |
|----------|------------------------|--------------------------|--------------|-------------------------------------|-------------------|
| A0048    | 12                     | 7.5                      | 16           | 25.5                                | 81                |
| A0026    | 12                     | 6                        | 20           | 24                                  | 83                |
| A0045    | 12                     | 7                        | 13           | 24                                  | 75                |
| A0036    | 12                     | 8                        | 12           | 24                                  | 75                |
| A0020    | 12                     | 7.5                      | 13           | 25.5                                | 77                |
| A0025    | 13.5                   | 8.5                      | 11           | 25.5                                | 78                |
| A0042    | 12                     | 7                        | 20           | 22.5                                | 82                |
| A0064    | 10.5                   | 8                        | 17.5         | 24                                  | 80                |

2.3. Analysis.

According to Kelley [17], reliability is the matter of how accurately a test measures a certain item that it does measure, while validity is the extent to which a test measures what it claims to measure. Furthermore, Guilford [18] stated that ‘a test is valid for anything with which it correlates’.

Cronbach’s α reliability coefficient normally ranges between 0 and 1.0. The closer Cronbach’s α coefficient is to 1.0, the greater the internal consistency of an item in the scale. According to George and Mallery [19], value > 0.9 is excellent, value > 0.8 is good, value > 0.7 is acceptable, value > 0.6 is questionable, value > 0.5 is poor, and value < 0.5 is unacceptable.

Observation on the Point Measure Correlation (PT-MEASURE) to detect polarity items is intended to examine the extent of the construction of constructs to achieve its goal. If the value contained in the PT-MEASURE CORR is positive (+), it indicates that the item measures the constructs to be measured [20]. However, if the value is negative (-), the item does not measure the constructs to be measured. Therefore, it needs to be improved because the item indicates that the outcome assessment implemented has a lack of validity.

3. Results

The Rasch analysis allowed the investigation on the reliability of the person (student) and item (attribute). Result of the analysis is shown in Tables 3 and 4.
The Cronbach’s α value of 0.79 revealed that the attribute was acceptable in assessing the students’ performance. Item reliability was at 0.98, which is an excellent reliability value [21]. This indicated that the assessment implemented had high reliability, thus appropriate and effective in measuring student performance. Person reliability was quite good at 0.81, which indicated a good spread of person with high levels of consistency in the students’ achievement.

Based on Table 5 below, all items were positive in PT-MEASURE CORR, which indicated that the items measured the constructs to be measured [20].

Table 5 shows that the five attributes met all the criteria as quality attributes because if the value of PT-MEASURE CORR is between 0.4 and 0.8, the item shall be accepted [22]. If the PT-MEASURE CORR is high, then the item is able to distinguish between students’ capabilities. Only one attribute did not meet the criteria, which is group work, because the value of PT-MEASURE CORR was less than 0.4.

Observation made from the map in Figure 2 indicated that five of the attributes located above the mean item and group work attribute were located below the mean. The group work score that came from peer assessment appeared to be the easiest attribute among those six attributes. Problem-solving skill appeared to be the hardest attribute among the attributes.
4. Discussion

This study investigated the validity and reliability of assessment methods through six attributes that have been implemented using learning outcome results of Year 4 students who took the Civil Engineering programme as a pilot study. The result was analysed using Winsteps software to perform a Rasch analysis. Overall, the test showed that the student assessments that have been implemented were valid and reliable. Furthermore, it demonstrated that these students performed well in their programme through the result of their learning outcomes.

There is a significant relationship between reliability and validity of assessment. An assessment that has high reliability is supposed to have high validity. Instead, a measurement with very poor reliability or validity may not be fit for its purpose. Therefore, validity and reliability are important when attempting to assess student performance in any education model implemented. An understanding of validity and reliability allows educators to make decisions about their education model that improves the performance of their students in terms of academic and personality.

There is a need to develop an instrument that can measure the implication of any education model implemented by the institutions to make sure their method can be used to measure student performance. There are several factors of an ideal measurement such as valid, reliable, and useful. Further research needs to be conducted to find the other factors and measure those factors such as by using the Rasch analysis to test the validity and reliability of the outcome assessment implemented.

References

[1] Miller M D, Linn R L and Gronlund N E 2013 Measurement and Assessment in Teaching 11th ed. Boston: Pearson

[2] Spady W G 1994 Outcome-based education: critical issues and answers, Arlington, VA: American Association of School Administrators

[3] Dunn K E and Mulvenon S W 2009 A critical review of research on formative assessments: the
limited scientific evidence of the impact of formative assessments in education. *Practical Assessment Research and Evaluation* vol 14(7)

[4] Klein S P, Kuh G D, Chun M, Hamilton L and Shavelson R 2005 An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education* vol 46 pp 251-276

[5] Timmerman B E 2008 Peer review in an undergraduate biology curriculum: effects on students’ scientific reasoning, writing and attitudes

[6] Hamid R, Baharom S, Hamzah N, Badaruzzaman W H W, Rahmat R A O K and RaihanTaha M 2012 Assessment of Psychomotor Domain in Materials Technology Laboratory Work. *Procedia - Social and Behavioral Sciences* vol 56 pp 718–723.

[7] Allen J D and Tanner K 2006 Rubrics: tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE-Life Sciences Education* vol 5(3) pp 197-203.

[8] Kumari N A, Rao D N and Reddy M S 2017 Indexing student performance with fuzzy logics evaluation in engineering education. *International Journal of Engineering Technology Science and Research* vol 4(9).

[9] Wu M and Adams R 2007 Applying the Rasch model to psycho-social measurement: A practical approach. *Melbourne: Educational Measurement Solutions*.

[10] Zain Z M 2018 Direct Assessment of Student Outcomes in Object Oriented Design Course Using Rasch Model Analysis. *Journal of Engineering Science and Technology* vol 13(12) pp 3908–3921.

[11] Ataei S, Mahmud Z and Khalid M N 2014 Identifying Potential Misfit Items in Cognitive Process of Learning Engineering Mathematics Based on Rasch Model. *Journal of Physics: Conference Series* vol 495(1) pp 12-26.

[12] Bichi A A, Talib R, Atan, N A, Ibrahim H and Yusof S M 2019 Validation of a developed university placement test using classical test theory and Rasch measurement approach. *International Journal of Advanced and Applied Sciences* vol 6 pp 22-29.

[13] Ismail A, Roslan L and Adnan A N 2017 Assessment on course outcome performance using Rasch measurement model. 2017 *IEEE 9th Int. Conf. on Engineering Education (ICEED) (Kanazawa)* pp 110-113.

[14] Khusaini N S, Saad N H, Aziz N, Ismail A and Kasolang S 2017 Relationship Between Demographic Attributes and Final Examination Performance Using Rasch Model. *Engineering Education (ICEED), IEEE 9th Int. Conf.* pp 117-121.

[15] Govindasamy P, del Carmen Salazar M, Lerner J and Green K E 2019 Assessing the Reliability of the Framework for Equitable and Effective Teaching With the Many-Facet Rasch Model. *Front. Psychol.* vol 10 p 1363.

[16] Razali S N and Shahbodin F 2016 Questionnaire on perception of online collaborative learning: Measuring validity and reliability using rasch model. *2016 4th Int. Conf. on User Science and Engineering (i-USEr) (Melaka)* pp 199-203.

[17] Kelley T L 1927 *Interpretation of educational measurements*. Yonkerson-Hudson, NY: World Book p 14.

[18] Guilford J P 1946 New standards for test evaluation. *Educational and Psychological Measurement* vol 6 pp 427-439.

[19] George D and Mallory P 2003 *SPSS for windows step by step: a simple guide and reference 11.0 update* (4th ed.) Boston: Allyn and Bacon.

[20] Bond T G and Fox C M 2015 *Applying the Rasch Model: Fundamentals Measurement in the Human Sciences* 3rd ed. New York: Routledge.

[21] Fisher W P 2007 Rating scale instrument quality criteria. *Rasch Measurement Trans.* vol 21(1) p 1095.

[22] Arsad N, Kamal N, Ayob A, Sarbani N, Tsuey C S, Misran N and Husain H 2013 Rasch Model Analysis on the Effectiveness of Early Evaluation Questions as a Benchmark for New Students Ability. *International Education Studies* vol 6(6) p 185