Assessing Students’ Learning Ability In A Postgraduate Statistical Course: A Rasch Analysis
Zamalia Mahmuda *, Nor Azura Md Ghani b, Rosli A. Rahim c

Abstract
In an effort to change the assessment paradigm from the traditional method of assessment, this paper will suggest a different assessment approach focusing on learning of statistics at the postgraduate level. As a case study, a course taught to Universiti Teknologi MARA postgraduate students in Applied Statistics, namely Categorical Data Analysis was chosen as the agent of assessment where students perceived ability based on an entrance-exit survey are assessed using Rasch analysis. This analysis is able to classify students’ perceived learning ability and identify learning difficulty more precisely. The study had shown that through Rasch measurement tools, students’ learning in a statistical course can be accurately measure based on the logit scale measurement derived from the Rasch model. This model has provided an excellent alternative tool for measuring postgraduate students’ actual ability in learning statistical concepts.

1. Introduction
A study on the teaching and learning of statistics at the postgraduate level was undertaken as part of a concerted effort to comply to Malaysia Quality Assurance (MQA) framework (MQA, 2012). MQA adopts the American Accreditation Board of Engineering and Technology 2000 (ABET, 2000) principles, which promote outcome, based education (OBE) learning process. OBE calls for the evaluation of the course learning outcomes (CLO) as specified in each course outline. Performance measurement has been largely dependent on students’ performance in carrying out task such as tests, quizzes and/or submission of assignments. Evaluation on the performance outputs categorized as mastery of knowledge and skill development, gives an indication on the achievement of the subjects’ expected CLO.

Over the past few years, there has been a concerted effort to improve the learning of statistics at the postgraduate level. The use of technology in the statistics classroom as well as new and innovative teaching strategies continue to offer students with many learning alternatives among them through e-learning materials. Some of these innovative learning strategies have also been used to improve student’s ability to learn better.

* Corresponding author name. Zamalia Mahmud. Tel.: +6-0122197985.
E-mail address: zamalia@tmsk.uitm.edu.my
However, assessment method has a big impact on how we diagnose students’ learning. It is acknowledged that current assessment could not completely identify learning problems. Therefore, a change of assessment paradigm is needed and this paper will demonstrate a different approach to assess students’ learning of statistics at the postgraduate level.

2. Study design and method

This study used a descriptive and survey method to gather information from twenty-five students who are currently pursuing their M.Sc. in Applied Statistics at Universiti Teknologi MARA, the largest local university in Malaysia. The survey is a section of the OBE procedure where students’ perceived learning are gauged through their responses of the entrance-exit survey. The study had adopted the Rasch measurement model (Bond & Fox, 2007; Wright & Stone, 1999) that can classify students’ learning abilities based on Anderson & Krathwohl (2001), a revision of Bloom’s Taxonomy knowledge dimensions. Results obtained were assessed using Rasch analysis where the strength of the measurement lies in its ability to calibrate between person ability to respond appropriately towards the survey items ($\beta_n$) and the level of difficulty of the survey items, ($\delta_i$), i.e. the agent of response.

Following the procedures of OBE, students’ perceived learning were gauged based on an entrance survey. The survey comprise of items/constructs developed based on the course learning outcomes. Entrance survey was administered at the beginning of the first class where students are required to respond to the entrance survey items in the e-learning portal environment, managed by the i-Learn centre at the university.

Students then went through the learning process in the following 14 weeks of lessons with coursework assigned to them and assessment given to numerically gauge their performance. At the end of 14th week, exit survey was administered to the students to measure their perceived learning after having gone through the lessons for the duration. Students are required to fill up the exit survey, which was displayed in the i-Learn portal.

2.1 Rasch Measurement Model

Rasch analysis is based on a stochastic or probabilistic model where Rasch measurement takes into account two parameters – test item difficulty and person ability. These parameters are assumed interdependent. However, separation between the two parameters is also assumed. For example, the items (questions) within a test are hierarchically ordered in terms of their difficulty and concurrently, persons are hierarchically ordered in terms of their ability. The separation is achieved by using a probabilistic approach in which a person’s raw score on a test is converted into a success-to-failure ratio and then into the logarithmic odds that the person will correctly answer the items. This is represented in a logit scale. When this is estimated for all persons, the logits can be plotted on one scale.

The items within the test can be treated in a similar manner by examining the proportion of items answered incorrectly and then converting this ratio into the logarithmic odds of the item being incorrectly answered. The logits can also be plotted on one scale. A person’s logit score can then be used as an estimate of that person’s ability and the item logit score can then be used as an estimate of that item’s difficulty. Since person ability was estimated from the proportion of correct answers and items difficulty from the proportion of persons with incorrect answers, both these estimates are related and the relationship between them can be expressed as a mathematical equation, i.e., Rasch measurement model as follows,

$$P_{ni} (x_{ni} = 1 / B_n - D_i) = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}}$$

(1)

The model expresses the probability of obtaining a correct answer (1 rather than 0) as a function of the size of the difference between the ability ($B$) of the person ($n$) and the difficulty ($D$) of the item ($i$). This Rasch model is used to calculate person abilities, to calculate item difficulties, and then to plot the person abilities and item difficulties on the same scale. According to the model, the probability of a person being successful on a given item is an exponential function of the difference between that person’s ability and the difficulty of the item. (Smith & Smith, 2004; Bond & Fox, 2007)
2.1.1. Assessing quality of data

A Rasch analysis is a procedure to assess the quality of raw score data using the Rasch model criteria such as fit statistics, z-standard residuals, and biserial correlations (Sick, 2010). A thorough Rasch analysis involves checking the degree to which the data match a unidimensional measurement model, identifying and diagnosing sources of discrepancy, removing items or persons if they are degrading the overall quality of measurement, and finally, constructing measures which, to the degree that the data approximate the Rasch model, are both interval level and sample independent. Infit and outfit mean square fit statistics provide summaries of the Rasch residuals, responses that differ from what is predicted by the Rasch model, for each item and person. High mean square fit statistics indicates a large number of unexpected responses. High person mean square values indicate test takers who filled in responses randomly, have unusual gaps in their knowledge. Item infit mean square values between 1.5 and 2.0 are considered to be unproductive for measurement, and values higher than 2.0 actually degrading (Wright & Linacre, 1994).

The Rasch model is not intended to fit data or to be evaluated by how well it fits any particular data set. The Rasch model is a definition of measurement derived from the universally accepted measurement requirements that: (i) The measures of objects be free of the particulars of the agents used to estimate these measures and the calibrations of agents be free of the particulars of the objects used to estimate these calibrations, (ii) The measures of objects and calibrations of agents function according to the rules of arithmetic on a common scale so that they can be analyzed statistically, and (iii) Linear combinations of measures and calibrations correspond to plausible concatenations of objects and agents.

Rasch expected that a valid instrument to have correct construct of linear scale which can be zero set and duly calibrated. A valid instrument can then be replicated for use and independent of the subject who used the instrument. Hence a reliable data is used for meaningful analysis and examination to generate useful information. This information is of utmost importance to be the prime ingredient in a particular decision making (Schau, Dauphinee, Stevens, & Del Vecchio, 1995; Wise, 1995). Based on Rasch measurement concept, correct response should be based on correct and truthful answer while inappropriate response could be due to response errors. These errors can be removed by focusing on the reproducibility of the latent trait measurement instead of forcing the expected generation of the same raw score, i.e. the common expectation on repeatability of results being a reliable test. The concept of reliability takes its rightful place in supporting validity rather than being in contentions. Hence, measuring learning ability in an appropriate way is vital to ensure valid quality information can be generated for meaningful use, i.e. by absorbing the error and representing a more accurate prediction based on a probabilistic model (Wright & Mok, 2004).

3. Analysis and results

3.1. Summary measure of reliability and separation index

In the analysis of entrance-exit survey, presents a summary measure about whether in general the data had shown acceptable fit to the model. The mean infit and outfit for person and item mean squares are close to 1.0. The mean standardized infit and outfit are between 0 to -.4. for persons and items. These measures are initial indication of data. This indicates the items are slightly overfit and that the data fit the model somewhat better than we would expect which may signal possible redundancy of items (Bond & Fox, 2007).

The standard deviation of the standardized infit is an index of overall mistfit for persons and items. Using 2.0 as a cut-off criterion, both persons (standardized infit standard deviation = .56) and items (standardized infit standard deviation = .34) show an overall acceptable fit.

Separation is the index of spread of the person positions or item positions. If separation is 1.0 or below, the items may not have sufficient breadth in position. For persons, separation is 2.22 for the data at hand (real), and 2.53 when the data have no misfit to the model (model) indicating there are three levels of person ability. The item separation index is 6.58 which indicate item difficulty can be separated into six levels. Person and item separation and reliability of separation assess instrument spread across the trait continuum. “Separation” measures the spread of both items and persons in standard error units. It can be thought of as the number of levels into which the sample of items and persons can be separated. For an instrument to be useful, separation should exceed 1.0, with higher values of separation representing greater spread of items and persons along a continuum. Separation determines...
reliability. Higher separation in concert with variance in person or item position yields higher reliability. The person separation reliability estimate for this data is 0.83 which indicate wider range ability. The item separation reliability estimate is 0.98 which indicate items are replicable for measuring similar traits.

Note that the mean for items is 0.0. The mean of the item logit position is always arbitrarily set at 0.0, similar to standard (z) score. The person mean is 0.21 which suggest that most items were well matched and easily endorsed or easy to agree with.

3.2. Person-Item Distribution Map (PIDM) of Entrance-Exit Survey

Figure 1 illustrates the Person-Item distribution map which is the heart of the Rasch analysis. Persons are distributed on the left side of the logit ruler (center vertical line) and items are distributed on the right side. “M” marks the person and item mean, “S” is one standard deviation away from the mean, and “T” is two standard deviation away from the mean. Those at the upper end of the scale agreed with more items and agreed more strongly (Bond & Fox, 2007).

In the map, we can see that about 98% of the students could not endorse the items as agreeable in the entrance survey, i.e., within two weeks of the first lesson. This can be seen where majority of the students on the left side of the logit ruler scale fall below most of the items on the right side of the logit scale. All these entrance items relate to their perceived ability to learn the concepts based on the course-learning outcome.

On the other hand, the reverse can be seen for the exit survey. About 98% of the students were able to endorse the items as agreeable, i.e. within one week after the completion of the course. This can be seen where majority of the students on the left side of the logit ruler scale are located above most of the items on the right side of the logit scale.

Figure 2 illustrates the Person-Item distribution map across the three course learning outcomes, namely CO1, CO2 and CO3. As in Table 1, CO1 was measured based on two learning outcomes (LOs), CO2 was measured based on thirteen LOs, and CO3 was measured based on two LOs. Within the component of CO2, about 99% of the students stated that they are not able to learn the course topics or concepts with the exception of one student (id 11711) which is considered as most able. Within the component of CO1 and CO3, only 50% of the students’ ability matched the range of difficulty of all three CO1 items but only two CO3 items.

At the exit survey, the map shows that all students could agree with the CO2 exit items and 99% could agree with all two CO3 exit items. Generally, about 98% of the students in the Categorical Data Analysis (CDA) class have perceived their ability to understand and perform about 95% of the CDA topics at the end of the course.

One item coded as CO2_8E, that is, all students perceived “I am able to compute Mantel Haenszel odds ratio and log odds ratio” as the most difficult topic. However, in the exit survey, CO2_8X was perceived as the easiest topic to understand by all students. This indicates students were able to learn Mantel Haenszel odds ratio topic extremely well.

Person-Item distribution map of person and entrance-exit survey items across the cognitive and psychomotor taxonomy levels. The map shows that students were having more difficulties in understanding concepts at the cognitive level than at the psychomotor level. At the entrance survey, 90% of the students perceived their lack of ability in understanding about 80% of the CDA topics at both cognitive and psychomotor levels. Item CO2_8En under the psychomotor category was perceived as most difficult. However, this situation changes at the exit survey where all students find the topic easy to learn. Topics which are categorized under the psychomotor level were perceived as topics which are easier to learn compare to topics categorized under the cognitive level.

The difference in students’ perceived ability between the entrance and exit survey is further illustrated using the empirical item characteristic curve (ICC). Figure 4 illustrates the empirical item characteristic curves for item CO2_8 at the psychomotor level and CO2_9 at the cognitive level. The logit measure on latent variable shows that at the entrance level, the distribution of logit scores were located at the bottom half of the Rasch logistic model curve, an indication of low person ability towards CO2_8 topic. The gap of approximately 6.9 logit scale between the entrance and exit item was an indication of the difference in the difficulty level. The situation, however changes at the exit survey where the distribution of logit scores were now located at the upper half of the logistic curve, an indication of a higher person ability and low item difficulty.
As for item CO2_9 at the cognitive level that is, “I am able to distinguish between different types of GLM”, the difference in the gap between entrance and exit item was about 4.6 logit scale, hence quite close compared to the gap in CO2_8 item.

4. Conclusion and Discussion

This paper had discussed the Rasch analysis approach in assessing students’ learning of statistics at the postgraduate level. This paper has demonstrated that students’ learning can be assessed in ways that will help the teacher to diagnose students’ learning problem better. As in the traditional assessment method, raw score is used to grade student according to their achievement. However, traditional assessment could not specifically measure the actual ability of students in grasping the difficulty of learning certain concept. Rasch analysis on the other hand, has the ability to diagnose and identify exactly where students are having difficulty the most in understanding certain concepts, and the extent of the learning difficulty that the student face in certain topic. The assessment tool used in this study to some extent coincide with previous studies which also discovered that students’ learning can be truly measured using Rasch measurement model (Noor Lide, Nor Zatul-Iffa, Zamalia & Mohamad Said, 2010; Cavanagh & Waugh, 2011).

Since this is one of the early diagnosis of identifying students’ learning problem, the assessment will continue to the next level where actual ability is assessed through students’ demonstration of actual practice questions and subjective tests.

Acknowledgements

We would like to thank the Ministry of Higher Education of Malaysia for the fund under the Fundamental Research Grant Scheme [600-RMI/SSP/FRGS 5/3/Fsp (40/2011)]. The authors also wish to express their gratitude to the Research Management Institute of Universiti Teknologi MARA for their undivided and continuous support in motivating the researchers.

References

Sick, J. (2010), Assumptions and requirements of Rasch measurement, JALT Testing & Evaluation SIG Newsletter, 14:2.