Sains Malaysiana 41(9)(2012): 1171–1176

Reliability Analisis of Examination Questions in a Mathematics Course Using Rasch Measurement Model

(Analisis Keupayaan Soalan Peperiksaan Kursus Matematik Menggunakan Model Pengukuran Rasch)

ZULKIFLI MOHD NOPIAH*, MOHD HELMI JAMALLUDDIN, NUR ARZILAH ISMAIL HALIZA OTHMAN, IZAMARLINA ASSHAARI & MOHD HANIF OSMAN

ABSTRACT

Students' performance cannot be based solely on their ability to answer questions. The accuracy of the examination questions must also be considered when measuring the outcome of the course. The objective of this paper was to study the reliability of a question and its impact on students' performance. A well-constructed question should be commensurate with the level of the intended knowledge. In this study, results from a linear algebra examination were analysed using the Rasch Model. The Rasch Model was used to analyse the reliability, non-redundancy and suitability of examination questions. The results showed that, in this particular case, the linear algebra questions were correctly constructed without any redundancy and suitable for the intended students.

Keywords: Bloom's taxonomy; item redundant; misfit item; person-item distribution; Rasch model

ABSTRAK

Prestasi pelajar tidak hanya bergantung kepada keupayaan pelajar menjawab soalan peperiksaan. Prestasi pelajar juga bergantung kepada kesesuaian soalan peperiksaan dalam mengukur hasil kursus pembelajaran. Oleh yang demikian, objektif kertas kerja ini ialah untuk mengenal pasti keupayaan soalan dan impaknya kepada prestasi pelajar. Soalan yang baik seharusnya merangkumi aspek tahap pengetahuan yang dikehendaki. Di dalam kajian ini, keputusan peperiksaan algebra linear dianalisis menggunakan Model Rasch. Model Rasch digunakan untuk mengkaji keupayaan, pertindihan dan kesesuaian soalan. Bagi kajian ini, hasil keputusan menunjukkan soalan algebra linear yang digunakan adalah bersesuaian tanpa menghasilkan pertindihan soalan dan memenuhi objektif pengajaran kursus tersebut.

Kata kunci: Item berulang; item ketidakpadanan; model Rasch; taburan person-item; taksonomi bloom

INTRODUCTION

The quality of education of any system depends on the input to the system, the performance within the system and the output from the system (Killen 2000). Recently, much attention has been focused on how to evaluate the return on investment in the public education system (outcome). Thus, a new set of theories (philosophy) of education that focus on systematic construction and evaluation of students' learning experiences have been introduced. This systematic approach to planning, delivering and evaluating instruction is called Outcome-Based Education (OBE). In the OBE system, two related outcomes are measured: the performance outcome and the intrinsic outcome. The performance outcome can be measured through test results and completion rates. The intrinsic outcome is expressed in terms of student ability, attitudes and personal traits.

Student achievement is a major concern for all parties, including universities, educators and parents. Certainly the university aims to produce students who are successful and are able to practice what they have learned to help their communities and country. Teaching and learning methods should therefore be based on the ability of the student to understand the subject that is being taught. Students' understanding of a subject is the most important aspect of learning and has long been supported by many researchers and educators (Sun et al. 2009). In the process of building understanding among students, educators and universities, relevant parties have to provide suitable assessment tools in their mode of teaching. Educators should provide a level of assessment commensurate with the student's cognitive level of thinking. Thus, one prerequisite for improving student performance is establishing a good examination.

A good examination must provide questions on what students have learned and be at the students' level of cognitive thinking. Felder et al. (2004) criticised some educators who blame students for poor achievement when, in actuality, the students have been given questions that are not at the same level as what they have been taught. Consequently, one cannot solely blame students when their performance are lower than expected. Improvements in student performance depend on whether educators can

adjust their questions to the level of the students' ability. Felder et al. (2004) states, "The best way to facilitate the development of higher-level of skills is to include high-level tasks in learning objectives, share them with the students in study guides for exams, give illustrations and practice in class and more practice on assignments".

The main question lies in how student performance is evaluated correctly. One of the most reliable and suitable methods of assessing student ability is using the Rasch Model. The Rasch Model is used to measure abilities, attitudes and personal traits from assessment data. Draugalis et al. (2004) applied the Rasch Model to evaluate student and item performance and assess curriculum strengths and weaknesses using a 65-item, multiple-choice examination. In another study, Azrilah et al. (2008) used the Rasch Model to validate the construct of measurement instruments. The Rasch measurement is found to give a better exploratory depth in understanding the expert level of agreement of an attitude. Zulkifli et al. (2010) applied a dichotomous Rasch model using 0-1 scoring of item response on multi-objective questions related to a linear algebra course at the Universiti Kebangsaan Malaysia. The study concluded that the Rasch Model is suitable for use in measuring both student ability and question validity.

This paper focused on measuring the reliability of exam questions on the specified objectives of a given course using the Rasch model. The idea behind this study was to enhance students' success with suitable sets of questions. There are times when educators create questions with very high standards when students are still relatively new to the subject. To construct questions that suit students' level of thinking, a correct analysis of the questions needs to be properly performed.

Method

The results of first-year students' examination papers in a linear algebra course from the Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia were considered in this study. A pool of data from a group of 215 students from different educational backgrounds was collected. Most were first-year students from the Faculty of Engineering and Built Environment. All data from the results were successfully extracted and screened into relevant tables. The data were normalised and then rated to the scale prior to the Rasch analysis.

Linear Algebra Results

The final examination questions of the linear algebra course consisted of three parts, Part A, Part B and Part C. The students were required to answers all questions in Parts A and B, whereas Part C was optional. There were 19 questions including the sub-questions considered in this study. Because there were different total marks for each question, the study used a standardisation method. The formula for the standardisation method is given below:

$$z_{ij} = \frac{x_{ij} - \min x_j}{\max x_j}, \quad (1)$$

where I is the i-th students (i=1,2.....215), j: j-th question (j=1,2....19), z_{ij}: standardised marks for i-th student and j-th question , x_{ij}: marks for i-th student and j-th question ($0 \le z_{ij} \le 1$), min x_j: minimum marks for j-th question ($0 \le x_{ij} \le 14$), max x_j: maximum marks for j-th question

The responses from the students' exam results were analysed using a rating scale in which the students were rated according to their achievement. From equation (1),

$$A_{ij} = z_{ij} \times 10. \quad (2)$$

Then, A_{ij} is classified into five groups of marks, such as 0-1.49, 150-3.49, 3.50-6.49, 6.50-8.49, 8.50-10 correspond to the rating scale '1', '2', '3', '4' and '5'. Scale 5 represents the highest marks that the students obtained. In the Rasch model, the probability of success can be estimated for the maximum likelihood of an event (Azrilah et al. 2007):

$$P(\theta) = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}},$$

where e is the base of natural algorithm or Euler's number; 2.7183, β_n is the student's ability, δ_i is the item or task difficulty, $P(\theta)$ is the learning ability.

The results were then tabulated and run in the Rasch Model software. In this paper, only two levels were specified: application (AP) and comprehension (CM), which are shown in the PIDM in Figure 1.

Results and Discussion

The summarised statistics of the result are given in Tables 1 and 2, which summarise the persons and items involved in this study. In the Rasch Model, persons represent the students, and the items represent the questions asked. The summary statistics contain information of the mean, standard deviation, and maximum and minimum values for both persons and items, where the maximum and minimum of the person and item spread are reflected in the standard deviation (SD) in the Person Item Distribution Map (PIDM) in Figure 1. This is called the distribution of the students, and the questions are based on the logit ruler. The questions were analysed using the Item Measure table to check for the validity of the item. The questions were compared to three rules using the Point Measure Correlation, Outfit MNSQ and Outfit ZSTD.

Table 1 shows the summary statistics of persons contain the information of Person Mean, μ_{person}, which is 0.18 logit. This is the major finding of the summary, where the value of the logit showed the performance of the students to be above the expected performance. Equally important is the value of the Reliability of Cronbach Alpha, which revealed a fair value of 0.68. The analysis identified two groups of student separation (G=1.70) with only 69.77% (N=150) of the students found to be "good" students and 30.23 (N=65) found to be "poor" students.

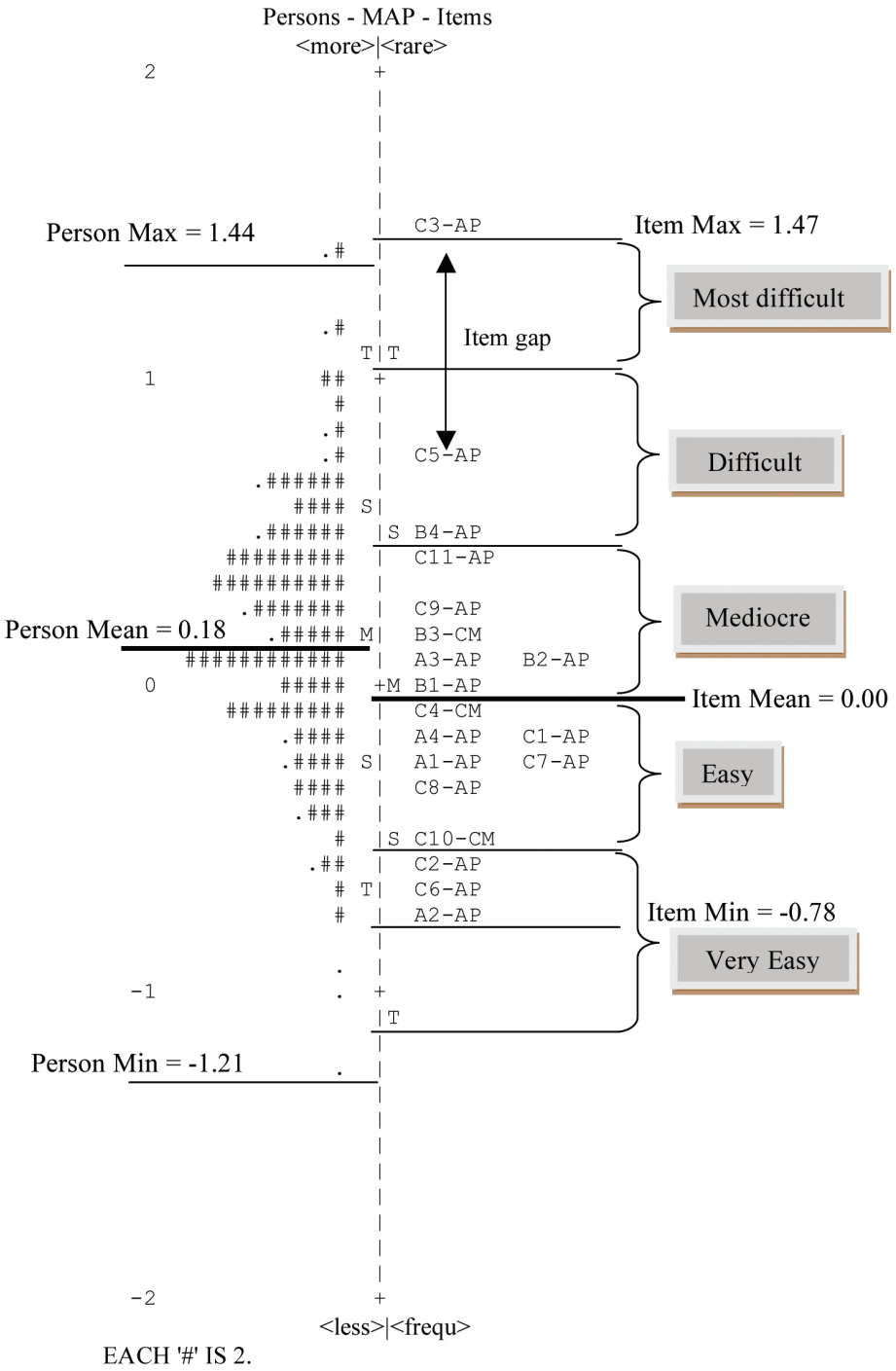

FIGURE 1. Person-Item Distribution Map (PIDM)

Regarding person reliability, the 0.74 value indicates a high consistency of a person to score either lower or higher than expected from the study.

The second set of summary statistics in Table 2 is the Items Summary. An item reliability of 0.97 indicates that the questions are reliable in measuring the proper item. An extension of the results can be viewed in Figure 1, the Person-Item Distribution Map (PIDM). The PIDM is a better picture of how the person correlates to the respective items and demonstrates better understanding, which shows a clearer view of the persons' ability and the relevant items' difficulty. The items' mean is used to establish the reference mark for the findings. The findings reveal that the students' performance far exceeded the expectation.

For this study, the items were the main focus. From the PIDM, all 19 questions were distributed on the logit scale. A higher ranking indicates that the item was more difficult and vice versa. The item labelled A3-AP was the

TABLE 1. Summary Statistics for persons

	Raw Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	51.1	15.1	0.18	0.21	1.01	0.0	1.01	0.0
Standard Deviation	11.1	0.4	0.44	0.04	0.31	0.9	0.45	0.9
Max	75.0	16.0	1.44	0.45	1.92	2.2	2.82	2.4
Min	20.0	12.0	-1.21	0.18	0.32	-3.1	0.23	-2.3
Real RMSE	0.22	Adj. SD	0.38	Separation	1.70	Person Reliability		0.74
Model RMSE	0.21	Adj. SD	0.39	Separation	1.85	Person Reliability		0.77

Valid responses: 79.3%

Person Raw Score-To-Measure Correlation = 0.98

Cronbach Alpha (KR-20) Person Raw Score Reliability = 0.68

TABLE 2. Summary Statistics for items

	Raw Score	Count	Measure	Model Error	Infit		Outfit	
					MNSQ	ZSTD	MNSQ	ZSTD
Mean	578.4	170.5	0.00	0.07	1.07	0.1	1.05	0.1
Standard Deviation	262.5	65.2	0.53	0.04	0.27	2.6	0.27	1.7
Max	968.0	215.0	1.47	0.20	1.58	3.3	1.88	2.7
Min	60.0	46.0	-0.78	0.05	0.54	-7.6	0.57	-5.2
Real RMSE	0.09	Adj. SD	0.52	Separation	5.57	Item Reliability		0.97
Model RMSE	0.08	Adj. SD	0.52	Separation	6.29	Item Reliability		0.98

UMean = 0.000 UScale = 1.000

Item Raw Score-To-Measure Correlation = -0.53

most difficult question, and the item labelled A2-AP was the easiest question. These are denoted by questions C3 and C5. The one question that students found particularly difficult to solve was question C3.

Examination questions (items) were classified into five different categories. The categories were divided based on the easy profiling of the items. As shown in Figure 1, the categories are 'most difficult', 'difficult', 'mediocre', 'easy' and 'very easy'. The spread of the items can be calculated using the difference between Item max and Item min, where 1.47 – (0.78) = 2.25 logit. The orthogonal arrow in the figure shows the gap between the two items. The wider the gap, the more difficulty the students encountered when attempting to answer the questions. Gap A3-AP is the largest gap, which means that the question of A3-AP was a difficult question.

Examination questions can overlap and be redundant. In the Rasch result, this can be detected under Item Dimensionality for each item, as shown in Table 3. A value below 0.7 indicates that no overlapping occurs. In this study, no overlapping of questions occurs, as indicated in the first column of Table 3.

The analysis of the Point Measure Correlation shown in Table 4 indicates whether the questions need to be further evaluated. The rules to determine a misfit item apply when all three rules have been violated and fall outside the range. These rules are the Point Measure Correlation, denoted by x where 0.4<x<0.8; the outfit mean square (MNSQ) as y where 0.5<y<1.5; and the outfit z-standard (ZSTD) as z where -2<z<2. For example, for item C5-AP the first rule shows the Point Measure Correlation of the C5-AP at 0.36, which falls outside the range of 0.4<x<0.8. This item has been categorised as a suspected misfit item.

For the second rule, the value of C5-AP for Outfit ZSTD = 2.6 falls outside the range and becomes a highly suspect misfit item. For the last rule, the value of C5-AP is bound within the range because the value of 1.48 falls within the range of 0.5<y<1.5. Therefore, C5-AP is no longer categorised as a misfit item. The same steps are used for all of the questions to determine if any of the questions is a misfit. The results show that no question in this study should be labelled a misfit item. Hence, all items are acceptable for further analysis.

CONCLUSION

Overall, this study showed that the questions used in the linear algebra examination are well-constructed. The results showed that the mean of students' performance is higher than the mean of the questions, which indicates that students are able to answer examination questions well within the given scope of the course. No overlapping

TABLE 3. Overlapping Items

Residual Correlation	Entry Number	Item	Entry Number	Item
0.33	5	B1-AP	6	B2-AP
0.30	16	C8-AP	17	C9-AP
0.27	14	C6-AP	15	C7-AP
0.23	18	C10-CM	19	C11-AP
0.22	10	C2-AP	12	C4-CM
0.21	9	C1-AP	10	C2-AP
-0.22	5	B1-AP	13	C5-AP
-0.22	7	B3-CM	9	C1-AP
-0.21	6	B2-AP	13	C5-AP
-0.21	1	A1-AP	5	B1-AP

TABLE 4. Item measure table

Entry Number	Total Score	Count	Measure	Model S.E	Infit		Outfit		Point Measure		Exact	Match	Item
					MNSQ	ZSTD	MNSQ	ZSTD	Corr	Exp	OBS%	EXP%	
11	60	46	1.47	0.20	1.33	0.9	0.95	0.2	0.37	0.40	78.3	81.0	C3-AP
13	370	192	0.78	0.06	1.39	3.3	1.48	2.6	0.36	0.45	26.6	33.2	C5-AP
8	527	215	0.47	0.05	1.10	1.3	1.12	1.1	0.48	0.50	19.5	23.5	B4-AP
19	482	190	0.44	0.05	1.19	2.2	1.05	0.5	0.54	0.50	18.9	22.5	C11-AP
17	551	190	0.26	0.05	1.11	1.4	1.03	0.3	0.52	0.51	15.8	19.2	C9-AP
7	670	215	0.14	0.05	0.96	-0.5	0.94	-0.6	0.55	0.52	17.2	19.9	B3-CM
3	677	215	0.12	0.05	0.54	-7.6	0.57	-5.2	0.52	0.52	42.8	20.1	A3-AP
6	679	215	0.12	0.05	1.17	2.2	1.11	1.1	0.56	0.52	11.6	20.1	B2-AP
5	733	215	-0.01	0.05	1.16	1.9	1.05	0.5	0.57	0.51	12.6	20.3	B1-AP
12	158	46	-0.11	0.11	1.58	3.1	1.88	2.7	0.44	0.54	13.0	25.9	C4-CM
4	785	215	-0.14	0.05	0.93	-0.8	1.19	1.6	0.37	0.50	31.2	23.5	A4-AP
9	161	46	-0.14	0.11	1.34	1.9	1.26	1.0	0.50	0.53	23.9	28.4	C1-AP
1	812	215	-0.21	0.05	0.75	-3.1	0.95	-0.4	0.35	0.49	25.6	27.0	A1-AP
15	750	192	-0.28	0.06	0.69	-3.5	0.77	-1.07	0.51	0.48	33.3	29.0	C7-AP
16	762	190	-0.34	0.06	0.82	-1.8	0.81	-1.3	0.52	0.46	40.0	33.1	C8-AP
18	804	190	-0.50	0.07	1.03	0.3	0.89	-0.6	0.52	0.43	49.5	43.6	C10-CM
10	195	46	-0.62	0.13	1.32	1.2	0.88	-0.2	0.53	0.44	43.5	42.6	C2-AP
14	846	192	-0.67	0.07	1.27	1.8	1.01	0.1	0/48	0.40	55.2	51.9	C6-AP
2	968	215	-.078	0.07	0.76	-1.6	0.91	-0.3	0.37	0.38	56.3	60.1	A2-AP
Mean	578.4	170.5	0.00	0.07	1.07	0.1	1.05	0.1			32.4	32.9	
Std Dev.	262.5	65.2	0.53	0.04	0.27	1.7	0.27	1.7			17.7	16.0	

questions have occurred in this study; thus, no questions must be revised, and no redundancy is apparent on the same topic. There is also no evidence of misfit questions. However, there is one concern from this study that requires further attention. The difficulty level of the questions needs to be revised. This is indicated by the gap between the most difficult question, A3-AP item, and the second most difficult question, C5-AP. The difference in logit between them is great. In summary, the Rasch Model provides a very useful tool to study the reliability of examination questions, and with its predictive feature, it is capable of overcoming the missing data. The use of the logit ruler is also useful for measuring specific outcomes, such as students' abilities and when validating a question construct online.

ACKNOWLEDGEMENT

The authors would like to express gratitude to Universiti Kebangsaan Malaysia (PTS 2011-021) for supporting the research. The author would also like to acknowledge the Centre for Engineering Education Research (p3k) and other supporting grants, PTS-2011-001 and UKM-OUP-NBT-28-135/2011, for the sponsorship of this journal's publication.

REFERENCES

Azrilah, A.A., Azlinah, M., Noor Habibah, A., Hamzah, A. G., Sohaimi, Z. & Saidfudin, M. 2008. Application of Rasch Model in Validating the Construct of Measurement Instrument. *International Journal of Education and Information Technologies* 2: 105-112.

Azrilah, A.A., Azlinah, M., Azami, Z., Sohaimi, Z., Hamzah, A.G. & Saidfudin, M. 2008. Evaluation of Information Professional Competency Face Validity Test Using Rasch Model. *5th WSEAS/IASME International Conference on Engineering Education (EE'08), Heraklion, Greece:* 396-403.

Azrilah, A.A., Azlinah, M., Noor Habibah, A., Sohaimi, Z. & Saidfudin, M. 2007. Appraisal of Course Learning Outcomes using Rasch Measurement: A Case Study in Information Technology Education. *International Journal of System Applications, Engineering & Development* 1: 164-172.

Draugalis, J.R. & Jackson, T.R. 2004. Objective Curricular Evaluation: Applying the Rasch Model to a Cumulative Examination. *American Journal of Pharmaceutical Education* 68 (2): 1-12. http://archive.ajpe.org/aj6802/aj680235/aj680235.pdf. [12 November 2011].

Felder, R.M. & Rebecca, B. 2004. The ABC's of Engineering Education: ABET, Bloom's Taxonomy, Cooperative Learning, and so on. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition.* http://www4.ncsu.edu/unity/lockers/users/f/felder/public/Papers/ASEE04(ABCs).pdf. [12 November 2011].

Killen, R. 2000. Outcomes-Based education: Principles and Possibilities. Unpublished manuscript, University of Newcastle, Faculty of Education. http://drjj.uitm.edu.my/DRJJ/CONFERENCE/UPSI/OBEKillen.pdf. [12 November 2011].

Sun, Y. & Chen, W. 2009. The Relationship Between Teaching Comprehensibility and Instructional Time Vs Students' Achievement in Rational Numbers. *The Journal of Human Resource and Adult Learning* 5(2): 99-107.

Zulkifli, M.N., Mohd Haniff, O., Noorhelyna, R., Fadiah Hirza, M.A. & Izamarlina, A. 2010. How good was the test set up? From Rasch Analysis Perspective. *Regional Conference on Engineering Education & Research in Higher Education (RCEE & RHEd)* Kuching, 7-9 June.

Centre for Engineering Education Research &
Unit of Fundamental Engineering Studies
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor
Malaysia

*Corresponding author; email: zmn@eng.ukm.my

Received: 25 May 2011
Accepted: 21 May 2012