Indian Journal of Science and Technology, Vol 8(12), DOI: 10.17485/ijst/2015/v8i12/70650, June 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Modeling a Multiple Choice Mathematics Test with the Rasch Model

Ahmad Zamri bin Khairani* **and Nordin bin Abd. Razak**

School of Educational Studies, Universiti Sains Malaysia, Penang, 11800, Malaysia; ahmadzamri@usm.my

Abstract

The purpose of the present study is to demonstrate adequacy of Rasch Model in modeling two important parameters in educational measurement, namely, students' ability and items' difficulty. 307 students provide response in a 40-item multiple choice test for the modeling using WINSTEPS 3.63. Results showed that 2 items do not fit the model's expectation, and thus dropped from further analysis. Statistics as well as positioning on the common scale for both parameters are also discussed. In short, the modeling is able to provide richer interpretations of the data collected.

Keywords: Item Difficulty, Mathematics, Multiple Choice Test, Rasch Model, Students Ability

1. Introduction

Multiple Choice Test (MCT) is one of the most widely use method for obtaining information regarding students' performance. This is because MCT provides several advantages such as ability to test quickly for large samples, automated scoring as well as providing quick feedbacks to both students and teachers[1,2]. In addition, MCT also provides better psychometric properties such as reliability and validity compared to other form of tests such as selected response tests (i.e. fill-in-the blank, essay) and performance tests (i.e. oral, presentation, lab work)[3]. Since reliability depends on number of items[4], MCT which normally consists of 30 to 50 items does have an advantage over 5-6 items essay test. Moreover, because the scoring of MCT is more objective compared to other forms, it demonstrates higher reliability measurements[4]. MCT also shows high evidence of content validity since it is able to cover more adequate samples of content[5]. Many guidelines for development of MCT are available with Haladyna and Downing[6] provide some exceptional reviews.

An important purpose of educational testing is to estimate students' ability in a particular subject so that important decision can be made. It is a common practice for schools to report the ability in terms of students' raw score, which is operationalized as the number of correct answer. High percentage of correct answers are associated with more able students whereas least able students will obtained the least number of correct answers. Nevertheless, the practice poses several shorcomings[7]. One important shortcoming is that, in interpretation using raw scores, it is assumed that a particular score represents the amount of ability. For example, if a student obtains 90% score, he or she is assumed to acquire the same amount of ability. However, as rightly observe by Wright and Masters[8] as well as Embretson and Reise[9], the assumption is rather skewed. Rather than having 90% ability, students who scores 90% simply means he or she has higher ability than most of the items adminstered in the test. Similarly, when a student fails to score any marks, it doesn't mean he or she has zero ability. It is more accurate to say that all items are more difficult for him/her.

2. The Rasch Model

In education, studies that address the abovementioned shortcomings of measurement are called test theories. There are two distinct test theories, namely, the Classical Test Theory (CTT) and the Item Response Theory (IRT). CTT deals mainly with test level scores such as raw score

*Author for correspondence

and reliability analysis whereas IRT focuses on item level statistics such as item difficulty and item discrimination. Literature shows that advantages offered by IRT over CTT are apparent, making it important for understanding measurement and testing today. IRT, also known as latent trait theory, relates responses of test items (observable trait) to students' ability (unobservable traits) using models that specify both traits[10]. Three IRT models have been developed. They are named for the number of parameters they use to estimate student ability. One parameter model, also known as the Rasch Model, uses only a single parameter, namely item difficulty to estimate unobservable trait (students' ability). Since it requires the least number of samples, it gives the Rasch Model a huge potential in terms of simplicity. Bond and Fox[11] give a comprehensive description of Rasch Model, both the model's principles as well as its applications. The two-parameter and three-parameter models are also widely used, especially in large scale assessment[12]. The two-parameter model adds an item discrimination parameter to the item difficulty, whereas the three parameter model adds a 'guessing' parameter to item difficulty and item discrimination. Substantial description of the two-parameter and three-parameter models as well as item response theory as a whole is available in the work of Keeves and Alagumalai[13].

Rasch modelling involves two important parameters, which are, (1) students' ability, and (2) item difficulty. Student's ability parameter is calculated based on the ratio of number of correct items to number of incorrect ones. This score is then transform into equal interval score (call '*measure*') using natural log (*ln*) in a procedure called calibration. Since equal interval measure provides ruler-like measurement of unobserved construct, it is essential in providing precision of interpretations of the parameters' measurement. Meanwhile, item difficulty parameter is defined as proportion of number of student who answers incorrectly over those who answer the item correctly. Rasch Model calibrations also transform item difficulty parameter into equal interval measure using the same algorithm. Both student's ability and item difficulty parameters (called 'measures') are describe in log-odd or '*logits*' unit. In addition, Rasch Model analysis is also able to calibrate both students' ability and item difficulty into one common scale. Higher ability students will be placed at the upper end, whereas lower ability students are placed at the lower end of the common scale. Difficult items are ordered into upper end of the scale while easy items are placed at the lower end. Bond and Fox[11] provide a mathematical equation to specify the relationship between both parameters, where, probability of answering correctly item i, P_i, is explained by ability of student n, β_n, and difficulty of item i, δ_i.

$$p_i = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \quad (1)$$

As mentioned before, although Rasch Model provides avenue for richer interpretations of data, the model is also subjected to strict assumptions. Two important assumptions that must be examined before the equal interval characteristics of measurement can be employed are (1) data must fit the model's expectations, and (2) the measured construct must be unidimensional. Analysis of fit helps detect discrepancies between the Rasch model expectation and the data collected[11]. Like any modelling procedures, Rasch Model analysis provides users with several goodness-of-fit indices. Two of the most widely used indices are the infit Mean-Square (MNSQ) and outfit MNSQ. Infit MNSQ, the inlier-sensitive, is more sensitive to the pattern of responses to item targeted on the student whereas outfit MNSQ, the outlier-sensitive, is more sensitive to responses to items with difficulty far from the person. Items with big infit and outfit MNSQs values are considered as 'misfitting' items. These items need to be eliminated from further analysis because they are measuring 'noise' and do not contribute meaningful to the measurement of the intended construct. Unidimensionality, on the other hands, assumes that all items measure a single ability. The assumption is very important because in practice, constructs are rarely strictly unidimensional. For example, the construct of student's mathematical ability may also include other constructs such as mathematical communication, that is, how students explain their answers. Rasch Model offers straightforward procedure to investigate dimensionality of a test. The Principal Component Analysis (PCA) of residuals procedure enables users to identify second factor that may become a threat to unidimensionality assumption. In this procedure, the first factor has been already removed, and the intention is to search for secondary factors. If the second factor provides useful information, the unidimensionality assumption is considered as under threat.

3. Methodology

The sample for this study consists of 307, 14 years-old students from public schools in the district of Seberang Perai Utara, Penang. The pools of items used are self-developed

based on the content specified in the Form 2 Mathematics Curriculum Specifications[14]. Content of the 40-items MCT covers three strands, namely, (1) number, (2) shape and space and (3) relations that are considered as important domains in students' mathematical ability construct. This study employs Rasch Model software, namely, WINSTEPS 3.63[15] to model both students'ability and item difficulty parameters. The measures were determined through iterative calibration of both parameters using the Joint Maximum Likelihood Estimation (JMLE) procedure. In WINSTEPS 3.63, the values of both infit and outfit MNSQ between 0.7-1.3 indicate that the assumption of model-data fit is fulfilled[11]. In addition, if the second factor from the PCA of residuals has the strength of 5 or more items, then the unidimensionality assumption is considered violated[15].

4. Results and Discussions

Table 1 shows five items of the MCT according to the fit order. Since the expected value of both infit and outfit MNSQ is 1.00, Item B30 measures 57% 'noise' from the actual construct. Since the values of outfit MNSQ for B20, B9 and B15 exceed the 0.7-1.3 guidelines, the items are candidates to be dropped from further analysis.

There are two primary sources for misfitting items, which is (1) bad items or (2) bad response. Bad item must be dropped from further analysis because it corrupts the measurement process. For bad response, two options are available – drop the entire responses for a particular sample, or exclude only problematic responses. Rasch Model analysis provides opportunity to examine both sources. In this study, investigation of bad responses is conducted first. Result from the analysis shows that fitting problems for B30, B38 and B20 are related to bad responses, while B9 and B15, are considered bad items. This is because exclusion of bad responses from items B30, B38 and B20 produce acceptable values of infit and outfit MNSQ. However, the same procedure does not work with items B9 and B15.

Table 2 shows the final analysis of 38-items MCT after excluding both bad items and bad responses. All items meet the assumption of model - data fit since the values of infit and outfit MNSQ are within the acceptable range of 0.7 to 1.3. The second factor extracted from the PCA of residuals has the strength of only 3 items. As such, the MCT also fulfill the requirement for meeting the unidimensionality assumption. Item B2 is the easiest (measure = –2.49 *logits*) while Item B30 is the hardest (measure = 2.34 *logits*). Further investigations using qualitative data with both students and teachers would provide explanation why these items are considered easiest and hardest. Equal interval characteristics of the item difficulty scale enable the following interpretations to be made: Item B20 (measure = 0.88 *logits*) is about 3 times more difficult than Item B1 (measure = 0.28 *logits*). Meanwhile, Item B20 is about twice less difficult than Item B12 (measure = 1.83 *logits*). Information about difficulty among items is essential in tailoring teaching and learning Mathematics. More explanation and exercise is dedicated for difficult items (or topics) while mastery of easy items is essential for progressing into a more difficult concepts and procedures.

Similar interpretations can also be made from calibration of students' ability statistics. The following Table 3 shows students' ability statistics for 5 students from the sample. Student #3 provides the most variation from the model's expectation based on his/her infit and outfit MNSQ values. However, unlike the item diffculty parameter, this student is not dropped from further analysis; rather, his or her responses should be investigated further to know the reasons for such discrepancy. Also, based on their measures, Student #4 is two times more able compared to Student #5.

Apart from item difficulty and students' ability descriptive statistics, Rasch Model analysis also provide mapping between both parameters as depicted in Figure 1.

Table 1. Items' difficulty statistics – initial analysis (N = 40)

Entry Number	Raw Score	Count	Measure	Model S.E.	Infit		Outfit		PTMEA	
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	Item
30	41	304	1.48	.17	1.12	1.0	1.57	3.4	A–.17	B30
9	54	307	1.15	.15	1.18	1.8	1.56	4.1	B–.24	B9
15	32	307	1.77	.19	1.09	.7	1.41	2.1	C–.11	B15
38	56	302	1.08	.15	1.11	1.2	1.30	2.5	D–.06	B38
20	70	306	.80	.14	1.11	1.5	1.29	2.9	E–.04	B20

Table 2. Items' difficulty statistics – final analysis (N = 38)

Entry Number	Raw Score	Count	Measure	Model S.E.	Infit		Outfit		PTMEA	
					MNSQ	ZSTD	MNSQ	ZSTD	CORR.	Item
1	104	303	.28	.13	1.19	4.1	1.24	3.8	–.06	B1
2	259	296	–2.49	.18	.90	–.8	.76	–1.6	.40	B2
3	220	306	–1.46	.13	.80	–3.6	.71	–4.3	.61	R1
4	249	301	–2.11	.16	.84	–1.7	.67	–3.0	.53	B4
5	200	306	–1.13	.13	.92	–1.8	.88	–2.2	.42	B5
6	189	307	–.95	.12	.78	–5.7	.75	–5.8	.63	B6
7	169	306	–.66	.12	.82	–5.3	.81	–5.3	.57	R2
8	97	306	.40	.13	1.04	.7	1.03	.5	.20	B8
9	DELETED									B9
10	186	304	–9.4	.12	.89	–2.6	.90	–2.2	.45	B10
11	137	306	–.20	.12	.95	–1.5	.93	–1.8	.37	R3
12	32	304	1.83	.19	1.02	.2	1.08	.5	.10	B12
13	159	307	–.51	.12	.85	–4.9	.83	–5.0	.53	R4
14	133	305	–.15	.12	1.03	.8	1.02	.4	.24	B14
15	DELETED									B15
16	121	306	.03	.12	1.11	2.8	1.16	3.1	.09	B16
17	132	307	–.13	.12	.92	–2.4	.90	–2.6	.41	R5
18	100	305	.35	.13	1.00	.0	1.05	.9	.24	B18
19	81	306	.67	.13	1.05	.8	1.15	1.8	.14	B19
20	70	306	.88	.14	1.13	1.7	1.30	3.0	–.02	B20
21	98	307	.39	.13	.96	–.7	.98	–.3	.31	B21
22	143	303	–.31	.12	.89	–3.4	.90	–2.8	.45	B22
23	123	302	–.03	.12	1.12	3.1	1.15	3.1	.09	B23
24	93	305	.46	.13	1.04	.8	1.04	.6	.20	R6
25	153	305	–.44	.12	.97	–.9	.98	–.7	.33	B25
26	91	304	.48	.13	1.12	2.2	1.18	2.4	.05	R7
27	152	303	–.45	.12	.89	–3.4	.87	–3.5	.46	R8
28	99	306	.36	.13	1.13	2.7	1.27	3.9	.01	B28
29	78	307	.73	.14	1.12	1.8	1.28	3.0	.01	B29
30	20	283	2.34	.23	1.02	.2	1.11	.5	.08	B30
31	94	307	.45	.13	1.10	1.9	1.18	2.4	.07	B31
32	77	305	.74	.14	1.02	.3	1.06	.7	.20	R9
33	100	303	.34	.13	1.20	3.9	1.28	4.2	–.07	B33
34	67	305	.94	.14	1.07	1.0	1.28	2.6	.05	B34
35	86	305	.58	.13	1.11	1.9	1.20	2.4	.05	B35
36	147	302	–.37	.12	.91	–3.0	.89	–3.0	.44	B36
37	154	304	–.45	.12	1.04	1.2	1.04	1.1	.23	R10
38	50	296	1.29	.16	1.09	.9	1.27	2.0	.00	B38
39	168	302	–.67	.12	.92	–2.4	.90	–2.6	.42	B39
40	129	303	–.10	.12	.91	–2.7	.91	–2.2	.42	B40
MEAN	12.53	303.8	.00	.13	1.00	–.4	1.02	–.2		
S.D.	54.0	4.3	.94	.02	.11	2.5	.18	2.8		

Table 3. Students' ability statistics

Entry Number	Total Score	Total Count	Measure	Model S.E.	Infit		Outfit		PT-Measure		Examinee
					MNSQ	ZSTD	MNSQ	ZSTD	Corr.	Exp.	
1	10	38	−1.19	.40	.91	−.41	.88	−.3	.45	.36	10
2	13	37	−.75	.37	.95	−.3	.92	−.3	.42	.37	11
3	13	37	−.64	.37	1.35	2.2	1.40	1.7	−.04	.34	12
4	13	37	−.71	.37	1.08	.6	1.14	.6	.29	.38	10
5	8	37	−1.40	.42	.93	−.2	1.87	2.1	.27	.31	12

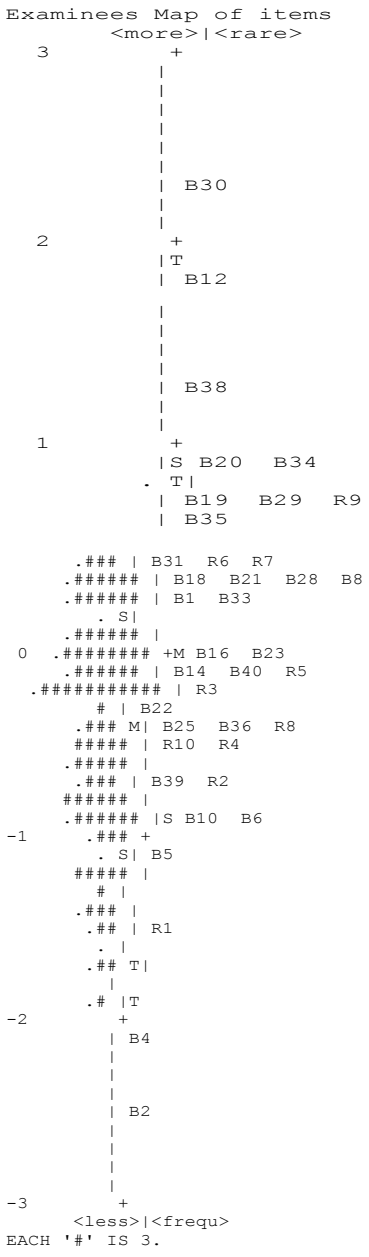

Figure 1. Mapping of students. Mathematical ability and items' difficulty.

The symbol '#' and '.' on the left hand side of the scale represents students' mathematical ability. Each '#' represents 3 students while each '.' represents one student. Students at the top of the represent more able students while least able students are place at the bottom. Meanwhile, on the right hand side, the symbol B1 to B40 shows the MCT items. Similar to the students' ability parameter, difficult items are placed at the top whereas easy items lie at the lower end of the scale. The letter M on the scale denotes means of both parameters. Higher mean of item difficulty over mean of students' ability means that this test is difficult for the sample of students. The information on mapping of parameters is essential for test developers. For example, in this MCT, in order to use the items in future, test developers should consider groups with higher ability students so that they will provide better estimations for those items. In addition, very easy items, correctly answered by everybody such as items B2 and B4 should be dropped from future use since the items do not provide much meaningful information about students' mathematical ability.

5. Conclusion

Richer interpretation of data through modeling using Rasch Model can be attributed to its ability to provide more statistics at item level compared to CTT. Statistics such as infit and outfit MNSQ provide indicators on quality of the measurement is. With regards to educationist, quality is essential since measurement provides information, which in turns provide basis of making decisions.

6. Acknowledgment

We gratefully acknowledge financial support from Universti Sains Malaysia under Research University grant 1001PGURU/81163.

7. References

1. Higgins E, Tatham L. Exploring the potentials of multiple-choice questions in assessment. Learning and Teaching in Action. 2003; 2(1):1–12.
2. Kuechler WL, Simkin M. How well do multiple choice tests evaluate student understanding in computer programming classes? J Inform Syst Educ. 2003; 14(2):389–400.
3. Wells SC, Wollack JA. An instructor's guide to understanding test reliability. Testing and Evaluation Services Publication. University of Wisconsin; 2003. Available from:http://www.wiscinfo.doit.wisc.edu/exams/Reliability.pdf
4. Miller MD, Linn RD, Gronlund NE. Measurement and assessment in teaching. Pearson Higher Education; 2012.
5. Popham WJ. Modern Educational measurement: Practical guidelines for educational leaders. Allyn and Bacon; 2000.
6. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ. 1989; 1:37–50.
7. Wright BD, Stone MH. Best test design. MESA Press; 1979.
8. Wright BD, Masters GN. Rating scale analysis. Embretson, SE: MESA Press; 1982.