ISSN 2229-5518

# Modelling Students' Mathematical Ability and Items' Difficulty Parameters: Application of the Rasch Measurement Model

Ahmad Zamri bin Khairani
School of Educational Studies, Universiti Sains Malaysia, 11800 Penang, Malaysia
ahmadzamri@usm.my
Nordin bin Abd. Razak
School of Educational Studies, Universiti Sains Malaysia, 11800 Penang, Malaysia
norazak@usm.my
Hasni binti Shamsuddin
Sekolah Menengah Sains Kepala Batas, 13200 Penang, Malaysia
emel_hasni@yahoo.com

**Abstract**— Measurement of students' ability is one of the most important purposes of educational measurement. Nevertheless, the purpose is considered difficult and inadequate based on the inappropriateness of the analyses used, especially when the students' ability measurement is always dependent of the test chosen for the studies. The purpose of this study is to explore the adequacy of the Rasch Measurement Model to provide so-called 'test-free' estimation of students' ability parameter based on their response in a set of items. A total of 411 Form 2 students were employed as sample for this study while a 40 multiple-choice Mathematics items provide a set of data for the modeling purpose. A Rasch Measurement Model software, the WINSTEPS 3.63 is employed for the purpose. Result showed that there is enough evidence of consistency between what been expected by the model and what been observed by the data. In short, results show that the Rasch Model analysis is able to provide richer interpretation towards better understanding of students' mathematical ability based on difficulty of the items. Implications of the results towards educational measurement are also reported.

**Index Terms**— *mathematical ability, item difficulty, students' ability, Rasch Measurement Model*

## 1 INTRODUCTION

One of the prime purposes in educational measurement is to estimate students' ability in a particular subject. Results from such a measurement will be used to make important decisions about the particular students. According to reference [1], this decisions includes (a) how to manage instruction such as planning instructional, monitoring students progress, diagnosing and assigning grades, (b) for counseling and guiding purposes, (c) placing students into different school's program, (d) selecting students for different purposes, and (e) for credentialing and certifying purposes. It is not uncommon that these decisions are very much influenced by students' ability. For example, in order to come out with a good planning of his or her teaching, teachers need to understand the ability of the students. High ability students should involve in activities that enhance their higher order thinking skills while lower ability student need more planning that will enhance their basic skills.

In reporting students' ability, most schools often use students' raw score, that is, the number of correct answer. A high percentage of correct answers are associated with more able student while lower student's score refers to a less able student. Nevertheless, the practice has several shortcomings. The most severe is that test scores are in ordinal, rather then interval scale as in a ruler. For explanation, although the test score can estimate the student's ability hierarchically, it cannot determine how this ability is different from the other student [2]. It can be shown that an increase of test score from 50 to 55 is not as easy as increases from 5 to 10. Also, it is not as difficult as to increase from 90 to 95. As such, test scores cannot distinguish accurately between the more able student and the less able one.

In addition, by using raw score to report students' ability, the assumption is that a raw score equals to the amount of ability of a student. One point of score is assumed to be equivalent to one unit of ability. For example, if a student scores 80% in a Mathematics test, he or she is assumed to have the same amount of ability. And, the student is considered to have twice the ability of another student who scores 40% on the same test. References [3] and [4] however, have demonstrated otherwise. One point of score is not equal to one unit of ability. A student who scores 80% does not mean he or she has acquired the same amount of ability. Similarly, students who score 0% do not mean he or she has no ability. In order to represent students' ability, the raw sores must be transformed into equal interval unit of measurement.

Besides inadequacy to represent students' ability, statistics obtained from raw sores such as p-value (the proportion of correct answer) are also sampled dependent. For example, a higher p-value will be obtained from a sample of above-

average student. Item with high p-value is considered an easy item. In contrast, below average sample will provide a lower p-value that indicates a more difficult item. Therefore, it can be seen that different interpretation can be made from the same single item. As a consequence, if the sample does not reflect the population, the item statistics obtained from the sample are limited in their usefulness. Similarly, since the raw score is defined in terms of number of correct answer, it is highly influenced by the test difficulty. Easier test will produce students with higher ability and vice versa. In short, since students' ability is test dependent, comparison among different students who sit for different tests does not provide a meaningful interpretation.

## 2. THE RASCH MODEL

In education, studies that study that address the shortcomings of measurement is called a test theory. The item response theory (IRT) is one of the most widely accepted test theories in educational measurement today. IRT relates responses of test items (observable trait) to students' ability (unobservable traits) through models that specify both traits [5]. Within the family of IRT, the Rasch Model is considered important in educational measurement based on several advantages. Unlike other IRT models, Rasch Model involves only one parameter, namely, the item difficulty, to estimate students' ability parameter; therefore, it is easier to work with. Secondly, in contrast with other models that accept all kind of data, Rasch Model provides users with element of choice where unwanted data such as guessing will not be entertained.

Like other IRT models, Rasch Model provides avenue to address the abovementioned measurement problems. In Rasch Model modeling, raw scores are transformed into equal interval 'measures' in a procedure called calibration where item difficulty parameter and student's ability parameter are estimated so that they can be put into a single scale. Student ability and item difficulty is measured using natural log and referred as log-odd unit or logits. Student's ability's parameter is defined as the number of correct items over number of incorrect one. For example, if an able student, n, correctly answers 20 out of 30 items, then the student's ability, bn is given by logits of,

$$\ln \frac{20}{10} = \ln 2 = +0.69$$

If a less able student, m, correctly answers 14 of the 30 items, then the student's ability, bm is given by logits,

$$\ln \frac{14}{16} = \ln 0.875 = -0.13$$

Student n is placed higher in the measured scale compared to student m. Item difficulty, on the other hand, is calculated as the number of student who answers incorrectly over those who answer that particular item correctly. For hard item, i, which is answered correctly by 56 out of 180 students, then item difficulty, δi, is given by logits of,

$$\ln \frac{(80-56)}{56} = \ln 2.214 = +0.79$$

For easy item, j, answered correctly by 120 of 180 students, then the item's difficulty, δj, is.

$$\ln \frac{(80-120)}{120} = \ln 0.5 = -0.69$$

Item i is placed at the upper end of the measured scale while item j constitutes the lower end. In summary, test calibration transforms raw scores into interval 'measures' in logits unit. Since the measures of both parameters are placed in a same scale, it permits direct comparison between students' ability and item difficulty.

The work of [3] provides a mathematical form to specify the relationship between both students' ability and items' difficulty parameters. Combining thedifference between students' ability and items difficulty enable researcher to explain the probability of student n response to item i. Since this difference, (βn – δi), vary from –∞ to +∞, applying the difference in terms of natural constant e = 2.71828 will limit the difference exp (βn – δi) between 0 and +∞. Furthermore, by taking the ratio of,

$$\frac{\exp(\beta_n - \delta_i)}{[1 + \exp(\beta_n - \delta_i)]}$$

the exponential expression of the difference would fit the probability value between 0 and 1. As such, the Rasch Model is represented by the following equation that specify a probability of a student n successfully answering an item i.

$$\frac{\exp(\beta_n - \delta_i)}{[1 + \exp(\beta_n - \delta_i)]}$$

Rasch Model has been used successfully used in various researches in education such as in test construction [6], [7], item and test analysis [8], and assessing psychometric properties of a test [9], [10]. The present study, however, seeks to provide empirical evidence on a more fundamental issue, namely, how well the data fits of the model's the expectation. This, in turns, provide better understanding of quality of the data. Unlike other IRT models, good data will provide good measurement of the construct while bad data is to be rejected since it will corrupt the measurement. In addition, the study also seek to examine how well the students' mathematical ability estimation concurs with the model's expectation

## 3. METHODOLOGY

The sample for the present study consists of 411 fourteen years-old students from public schools in the district of Seberang Perai Utara, Penang. Meanwhile, the pools of items used are self-developed based on the content specified in the Form 2 Mathematics Curriculum Specifications [11]. The test is hypothesized to measure mathematical ability construct which is conceptualized of having 3 sub-dimensions, namely, concep-

tual understanding, procedural fluency, and strategic competence (problem solving [12]. With regards to the data analysis, this study employs Rasch Model software, namely, WINSTEPS version 3.63 [13] to model both students' ability and item difficulty parameters. In WINSTEPS, the measures reported in 'logits' were determined through iterative calibration of both parameters using the Joint Maximum Likelihood Estimation (JMLE). WINSTEPS provide various statistics to provide evidence whether the data fits the model expectations. This study discusses two of the fit statistics, namely, the infit and outfit mean-square (MNSQ) and percentages of exact match between observation from the data and expectation from the model.

Infit MNSQ, the inlier-sensitive, is more sensitive to the pattern of responses to item targeted on the student, and vice versa. Outfit MNSQ, the outlier-sensitive, is more sensitive to responses to items with difficulty far from the person and vice versa. According to reference [14], if the behavior of the test has yet to be obtained, MNSQ values between 0.7 – 1.3 for every item is considered reasonable. Misfitting item shows the possibility of that particular item not being able to measure the same construct. It is also considered as a "weak" item that can influence test reliability. These responses need to be eliminated from further analysis because they are measuring 'noise' and do not contribute to the measurement of the intended construct. In short, fit statistics help test developer to decide upon the appropriateness of the items [15]. Similarly, the percentage of exact match between observation and expectation from the model shows whether the data are more random or more predictable that what the model predicts. The ideal result is for both percentages to be equal.

## 4. Findings and Discussions

Based on infit and outfit MNSQ statistics in Table 1, all items are within the acceptable range of 0.7 – 1.3. Meanwhile 13 items (32.5%) show exact match between the observation and expectation, while another 13 items are more random than the model predicts. In contrast, 14 items (35%) are more predictable. The finding is not unexpected because the present study employs relatively small sample. By increasing the number of sample, the model will be able to provide better prediction of the data. Nevertheless, since the variation is small between 0 .2% (Item 37) and 8.4% (Item 30), these items are considered productive for a measurement purpose. In short, both statistics show that there is enough evidence that the data obtained fits the model expectation. With regards to students' mathematical ability, 335 students (81.5%) show responses that are within the expectation of the model. The results give suggestion that the sample has contributed usefully to the measurement of mathematical ability construct.

## Table 1: Item Statistics according to Difficulty

| ENTRY NUMBER | TOTAL SCORE | COUNT | MEASURE | MODEL S.E. | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PT-MEASURE CORR. | PT-MEASURE EXP. | EXACT MATCH OBS% | EXACT MATCH EXP% | item |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 68 | 410 | 2.53 | .14 | 1.00 | .1 | 1.01 | .1 | .32 | .33 | 84.9 | 84.4 | Q5 |
| 40 | 79 | 411 | 2.32 | .13 | 1.08 | 1.0 | 1.11 | 1.0 | .25 | .34 | 80.8 | 82.0 | Q40 |
| 16 | 103 | 409 | 1.92 | .12 | .98 | -.3 | 1.01 | .1 | .36 | .35 | 78.5 | 76.9 | Q16 |
| 33 | 109 | 409 | 1.84 | .12 | 1.23 | 3.6 | 1.30 | 3.4 | .09 | .35 | 69.9 | 75.7 | Q33 |
| 4 | 116 | 410 | 1.75 | .12 | 1.08 | 1.4 | 1.17 | 2.1 | .25 | .35 | 73.7 | 74.5 | Q4 |
| 12 | 128 | 410 | 1.58 | .11 | .94 | -1.2 | .91 | -1.3 | .43 | .35 | 75.1 | 72.4 | Q12 |
| 30 | 131 | 407 | 1.54 | .11 | .87 | -2.7 | .88 | -1.9 | .49 | .36 | 80.1 | 71.7 | Q30 |
| 15 | 134 | 408 | 1.50 | .11 | 1.01 | .3 | 1.03 | .6 | .33 | .35 | 69.4 | 71.3 | Q15 |
| 38 | 154 | 410 | 1.27 | .11 | 1.11 | 2.5 | 1.16 | 2.8 | .22 | .35 | 63.4 | 68.5 | Q38 |
| 8 | 163 | 409 | 1.15 | .11 | 1.06 | 1.4 | 1.08 | 1.5 | .29 | .35 | 66.0 | 67.5 | Q8 |
| 29 | 167 | 408 | 1.10 | .11 | .87 | -3.6 | .84 | -3.3 | .50 | .35 | 74.8 | 67.0 | Q29 |
| 32 | 191 | 407 | .82 | .11 | 1.01 | .2 | 1.01 | .2 | .34 | .35 | 65.8 | 65.0 | Q32 |
| 19 | 201 | 409 | .73 | .11 | .99 | -.2 | 1.01 | .2 | .35 | .34 | 66.0 | 64.4 | Q19 |
| 6 | 209 | 406 | .62 | .11 | 1.02 | .5 | 1.02 | .5 | .32 | .34 | 64.8 | 64.0 | Q6 |
| 34 | 223 | 411 | .49 | .11 | .90 | -3.1 | .88 | -2.4 | .45 | .34 | 69.6 | 64.0 | Q34 |
| 31 | 231 | 411 | .40 | .11 | .85 | -4.8 | .80 | -3.8 | .51 | .33 | 72.0 | 64.2 | Q31 |
| 14 | 238 | 408 | .31 | .11 | 1.09 | 2.5 | 1.10 | 1.7 | .23 | .33 | 59.8 | 64.8 | Q14 |
| 24 | 248 | 410 | .20 | .11 | .97 | -1.0 | .93 | -1.2 | .37 | .32 | 66.8 | 65.5 | Q24 |
| 26 | 256 | 409 | .11 | .11 | .96 | -1.2 | .91 | -1.3 | .37 | .32 | 67.7 | 66.4 | Q26 |
| 10 | 261 | 406 | .02 | .11 | .99 | -.2 | .94 | -.9 | .33 | .31 | 62.8 | 67.3 | Q10 |
| 28 | 271 | 411 | -.06 | .11 | .99 | -.2 | 1.00 | .0 | .31 | .31 | 70.3 | 68.2 | Q28 |
| 36 | 278 | 411 | -.15 | .11 | .99 | -.2 | .93 | -.9 | .33 | .30 | 68.4 | 69.4 | Q36 |
| 37 | 284 | 411 | -.22 | .11 | 1.10 | 2.2 | 1.16 | 1.9 | .17 | .30 | 70.3 | 70.5 | Q37 |
| 9 | 284 | 410 | -.23 | .11 | 1.10 | 2.2 | 1.22 | 2.5 | .16 | .30 | 68.5 | 70.6 | Q9 |
| 7 | 297 | 410 | -.40 | .12 | 1.00 | .1 | 1.09 | .9 | .26 | .29 | 74.4 | 73.2 | Q7 |
| 21 | 305 | 410 | -.50 | .12 | 1.01 | .2 | .98 | -.2 | .27 | .28 | 74.9 | 74.9 | Q21 |
| 25 | 315 | 411 | -.64 | .12 | .93 | -1.1 | .85 | -1.5 | .35 | .27 | 77.9 | 76.9 | Q25 |
| 27 | 319 | 409 | -.73 | .12 | 1.00 | .0 | 1.01 | .1 | .26 | .26 | 78.5 | 78.2 | Q27 |
| 20 | 322 | 410 | -.76 | .13 | .98 | -.3 | .91 | -.8 | .29 | .26 | 78.0 | 78.7 | Q20 |
| 13 | 340 | 411 | -1.05 | .14 | .98 | -.2 | .92 | -.5 | .27 | .24 | 82.7 | 82.7 | Q13 |
| 2 | 351 | 411 | -1.26 | .14 | 1.07 | .7 | 1.16 | 1.0 | .12 | .22 | 85.4 | 85.4 | Q2 |
| 18 | 355 | 411 | -1.35 | .15 | .93 | -.7 | .74 | -1.7 | .33 | .21 | 86.4 | 86.4 | Q18 |
| 17 | 356 | 411 | -1.37 | .15 | .95 | -.5 | .81 | -1.1 | .29 | .21 | 86.6 | 86.6 | Q17 |
| 11 | 360 | 411 | -1.46 | .15 | .97 | -.3 | .94 | -.3 | .24 | .21 | 87.6 | 87.6 | Q11 |
| 35 | 365 | 410 | -1.61 | .16 | 1.04 | .4 | 1.18 | 1.0 | .13 | .20 | 89.0 | 89.0 | Q35 |
| 3 | 364 | 408 | -1.62 | .16 | .94 | -.4 | .79 | -1.1 | .28 | .19 | 89.2 | 89.2 | Q3 |
| 1 | 375 | 411 | -1.87 | .18 | 1.00 | .0 | 1.13 | .7 | .16 | .18 | 91.2 | 91.2 | Q1 |
| 39 | 383 | 410 | -2.19 | .20 | .96 | -.2 | .78 | -.9 | .23 | .15 | 93.4 | 93.4 | Q39 |
| 22 | 388 | 411 | -2.37 | .22 | .99 | .0 | .80 | -.7 | .18 | .14 | 94.4 | 94.4 | Q22 |
| 23 | 388 | 411 | -2.37 | .22 | 1.01 | .1 | 1.19 | .8 | .12 | .14 | 94.4 | 94.4 | Q23 |
| MEAN | 252.7 | 409.7 | .00 | .13 | 1.00 | -.1 | .99 | -.1 | | | 76.3 | 76.0 | |
| S.D. | 96.2 | 1.5 | 1.33 | .03 | .07 | 1.6 | .14 | 1.5 | | | 9.6 | 9.6 | |

Since the measures are in interval scale, one important observation that can be made from the finding is that the most difficult item, Item 5 (2.53 logits) is twice as difficult compared to Item 38 (1.27 logits). Similarly, Item 6 (.62 logits) is considered twice as easy compared to Item 38. Another important observation is that a bulk of difficult items consist of both algebra and connection item (where students need to connect two or more knowledge, skills and abilities) while easier items mainly consist of arithmetic items. As such, students with high mathematical ability can be operationally defined as to be able to master content related to algebra as well as to connect previously learned knowledge, skills and abilities to solve new problems. On the other hands, students with lower mathematical ability can only solve problems related to arithmetic. This definition would certainly helpful to provide standards for teachers to improve mathematical ability of the students. In summary, Rasch Model provides avenue for teachers and researchers to provide richer interpretations on the relationship between student's ability and test items compared to the traditional test theory.

## 5. ACKNOWLEDMENTS

This article is made possible by the funding obtained from the Universiti Sains Malaysia under Short Term Grant 304/PGURU/6311048

# 5. References

[1] A. J. Nitko, Educational assessment of students (2nd ed.), Merrill, Englewood Cliffs, NJ (1996)

[2] T. G. Bond, and C. M. Fox, Applying the Rasch model: Fundamental measurement in the human sciences, Lawrence Erlbaum, Mahwah, NJ (2001)

[3] B. D. Wright, and M. H. Stone, Best Test Design, MESA Press, Chicago, IL (1979)

[4] [4] B. D. Wright and G. N. Masters, Rating scale analysis, MESA Press, Chicago, IL (1982)

[5] S. E. Embretson and S. P. Reise, Item response theory for psychologists, Mahwah, NJ: Lawrence Erlbaum (2000)

**[6]** T. Forkmann, M. Boecker, N. Wirtz, M. Eberle, P. Westhofen, K. Schauerte, K. Mischke, and C. Norra., Development and Validation of the Rasch-based Depression Screening (DESC) using Rasch Analysis and Structural Equation Modeling, 3, 40 (2009)

[7] K. Y. Chang, M. Y. Tsou, K. H. Chan, S. H. Chang, J. J. Tai and H. H. Chen, Item Analysis for the Written Test of Taiwanese Board Certification Examination in Anesthesiology using the Rasch Model, British Journal of Anesthesia, 6, 104 (2010)

[8]

[9] P. Baghaei, A comparison of three polychotomous Rasch models for super-item analysis, Psychological Test and Assessment Modeling, 3, 52 (2010)

[10]

[11] K. R. Muis, P. H. Winne, and O. V. Edwards, Modern Psychometrics for Assessing Achievement Goal Orientation: A Rasch analysis. British Journal of Educational Psychology, 3, 79 (2009)

[12] K. Ahmad Zamri, Application of the Bookmark Method in Setting Performance Standards for Form 2 Students in Mathematics, unpublished doctoral thesis (2010)

[13] Curriculum Development Centre, Ministry of Education, Curriculum specifications for Mathematics Form 2 (2002)

[14] J. Kilpatrick, J. Swafford, and B. Findell, Adding it Up: Helping How Children Learn Mathematics, Washington DC: National Academy Press (2001)

[15] J. M. Linacre, A user's guide to Winsteps (2005)

[16] T. G. Bond, and C. M. Fox, Applying the Rasch Model: Fundamental Measurement in Human Sciences, Lawrence Erlbaum: Mahwah, NJ (2001)

[17] R. K. Green and C. G. Frantom, Survey development and validation with Rasch Model. http://www.ipsm.umd.ed/qdet/final_papers/green.pdf, retrieved April 23 (2012)