Advance in Educational Measurement: A Rasch Model Analysis of Mathematics Proficiency Test

Ahmad Zamri bin Khairani and Nordin bin Abd. Razak

***Abstract*—The purpose of this paper is to provide evidence of the adequacy of Rasch Model analysis in providing richer interpretation regarding examinee performance. The discussion centers on measurement of Mathematics proficiency construct using the analysis of a self-developed Mathematics Proficiency Test (MPT) with a sample of 588 14 years-old examinees. The result showed not only measurement of the construct met the Rasch Model assumptions; the test also demonstrated good psychometric properties**

***Index Terms*—Examinee proficiency item difficulty, mathematics proficiency construct, rasch model.**

I. Introduction

Scholars agree that the definition of the measurement always involves assigning numerical numbers to certain traits or characteristics using a tool [1]. For physical traits, such as height, the process of assigning numbers can be done directly using a ruler. However, psychological traits such as ability or proficiency are constructs. They are unobservable but can be measured indirectly using a tool called test. The design of tests to measure constructs, however, presents several problems. For example, no single approach to the measurement of any construct is universally accepted. Since the measurement of psychological constructs is always done indirectly, there is always the possibility that researchers will select different types of behavior to measure the same construct. As a consequence, different inferences will be concluded. Lack of well-defined units in the measurement scale also poses problem. For example, an examinee who is unable to answer any test item does not mean that he or she has "zero" ability. Instead, all the items have difficulty index which is more than the examinee's ability. The study of measurement problems and methods to overcome them is known as test theory. Test theories relate observable traits (such as *test score*) with unobservable traits (such as *ability* or *proficiency*) for a measured construct using mathematics model [2].

The first established test theory is called the Classical Test Theory (CTT). The CTT revolves around concepts of true score, measurement error and index of test reliability CTT relates observable trait (the test score, X) with the unobservable trait (the person's true ability on the characteristics, T) with the following equation:

$X = T + E$, where E = measurement error. Item Response Theory (IRT), meanwhile, relates responses to test items (observable trait) to unobservable traits through models that specify how both trait level and item properties are related to person's item response [3]. Three IRT models have been developed. They are named for the number of parameters they use to estimate examinee ability. One parameter model, also known as the Rasch Model, uses only single parameter, namely item difficulty to estimate an unobservable trait of a particular examinee. The two-parameter and three-parameter models are also widely used, especially in large scale assessment [4]. The two-parameter adds an item discrimination parameter to the item difficulty, whereas the three parameter model adds a 'guessing' parameter to item difficulty and item discrimination. [5] provide substantial description of the two-parameter and three-parameter models as well as item response theory as a whole.

One of the major limitations of the CTT is that the item statistics (the difficulty index, *p*-value) and (the discrimination index, *r*-values) which are very essential in the application of CTT are sampled dependent. These limitations are addressed and overcome in IRT. When its assumptions have been satisfied, IRT provides (1) examinee ability measures that are independent on the particular sample of test items chosen, (2) item statistics that are independent of sample of examinee drawn, and (3) fit statistics indicating the precision of the estimated ability for each examinee and precision of each item. A classic article by [6] provides readers with detailed explanation to invariant person and item parameter known as '*examinee-free*' test calibration and '*item-free*' examinee measurement. With ability estimates being invariant; IRT provides a way of comparing examinee even though they take a different test.

All IRT models offer invariant properties for estimation of item and examinee parameters. According to [7], the choice of appropriate model depends on the type of test questions and their scoring. Another important consideration is that, in practice, the choice of models depends on the amount of data available. The larger the number of parameter is, the more data are needed for parameter estimation, thus requiring more complex calculation and interpretation. In this case, Rasch Model has some special properties that make it attractive to users. Rasch Model involves fewest parameters; therefore, it is easier to work with [4]. [8] gives more influential explanation in favor of Rasch Model compared to a three-parameter model. These two models are opposite in philosophy and in practice. The three-parameter model will adjust to adapt whatever type of data (includes invalid responses). The Rasch model however has tight standards in controlling the data. Unlike the three-parameter model, invalid responses such as guessing on item will not be accepted. It is described as unreliable person reliability. Critics of the Rasch Model often regard the model as having strong assumptions that are difficult to meet.

Manuscript received March 1, 2012; revised April 4, 2012

This article is made possible by the funding obtained from the Universiti Sains Malaysia RU Grant 1001/PGURU/811163

Authors are with the School of Educational Studies, 11800 Universiti Sains Malaysia, Penang, Malaysia (e-mail: ahmadzamri@usm.my; norazak@ usm.my).

However, these are values that make Rasch Model more appropriate in practice.

One major problem in measurement lies in the interaction between the person being measured and the instrument involved. Performance of a person is known to be dependent on which instrument is used to measure his or her trait. However, this shortcoming is circumvented by procedure of conjoint measurement in Rasch Model. [9] explain that in conjoint measurement, the unit of measurement is not the examinee or the item, but rather the performance of an examinee relative to a particular item. If β_n is an index for ability for examinee *n* on the trait being measured, and if δ_i is an index for the difficulty of the item *i* which relates to the trait being measured, then the unit of measurement is neither β_n nor δ_i but rather $(\beta_n - \delta_i)$, which is the difference between the ability of the examinee and the difficulty of the item. If the ability exceeds the item difficulty, then it is expected that the examinee will answer the item correctly. In contrast, if the difficulty exceeds the ability, then it is expected that the examinee will answer incorrectly. In education, response on a particular item is always in uncertainties. Therefore, probabilistic approach has to be employed when explaining what happens when an examinee takes an item. Probabilities of correct response are between 0 and 1 and it does not permit proportion of correct answer to be expressed in interval scale. To overcome these constraints, logistic transformation, which involves taking the natural logarithm, is used. As a final product, it can be shown (that the probability of person *n* has correct response to item *i* is given by $P_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}$ [9]

Rasch Model offers procedure to transform test score into interval-scale measure (score) in log-odd or *logits* unit. Earlier works clearly showed a need for interval-scaled measures in measurement of an intended construct [10]. Application of summated score (such as number of correct answers) have been strongly opposed due to the fact that it is highly unlikely that examinee score can be interpreted accurately, cannot determine how one's score is different from other examinee and that the difference between two scores is not reliable when different scoring scheme is in used [11]. In Rasch Model analysis, two important parameters usually discussed are item difficulty and examinee ability. Item difficulty measure is an estimate of an item's underlying difficulty calculated from the number of examinee who succeeds in that item. Examinee's ability measure, on the other hand, is an estimate of his or her underlying ability based on performance on a set of items.

In order for the Rasch Model measurement to have the '*examinee-free*' item difficulty and '*item-free*' examinee ability measurement, two important assumptions must be met. Firstly, the data must meet the unidimensionality assumption, that is, they represent a single construct [11]. Secondly, Rasch Model requires that the data must fit the model [11]. In addition, it is also imperative to provide evidence on the psychometric properties of the test used from the framework of Rasch Model analysis. In light of the preceding discussion, the present study is aimed at (1) examine the extent to which a set of test to measure Mathematics proficiency meets Rasch Model expectation, and (2) provide evidence of adequate psychometric properties of the test.

II. Methodology

The sample for this study consists of 588 14-years old students (Grade 8) from the district of Lower Perak, Perak. The instrument used in this study is a self-developed 50-item Mathematics Proficiency Test (MPT). The MPT is developed according to guidelines provided in the Curriculum Specifications [12] and further validated by 3 experienced teachers. In this study, a Rasch Model software, WINSTEPS version 3.57 [13] is used. In WINSTEPS, the measures are determined through iterative calibration of both person and item using Joint Maximum Likelihood Estimation (JMLE). In WINSTEPS, the principal component analysis (PCA) of the residuals procedure helps identify the existence of second factor that pose a threat to unidimensionality assumption. The second factor of the strength of 3 items is considered not a threat [13]. On the other hands, the infit mean square (MNSQ) and outfit MNSQ provide indications of the discrepancies between the data and model's expectations. This study adopts the range of acceptable fit between 0.7 – 1.3 for both fit indices as suggested by [11]. Psychometric properties of the test were tested in terms of reliability and validity of the measures' meaning and interpretation. Rasch analysis provides reliability indices for both item and examinee's measure. High reliability for both indices are desirable since they indicate a good replication if the comparable items/examinees are employed.

Another analysis conducted is related to construct validity of the measures meaning and interpretation. Based on the foundation laid by [14] two major threats to construct validity that are under investigation are construct-irrelevant variance and construct under-representation. The former relates to the irrelevant variances that contaminate measurement of the main construct while in the latter, the measurement fails to include important sub-dimensions of the construct. In short, construct validity requires nothing irrelevant be added while at the same time nothing important should be left out in assessing a construct. Within the framework of Rash Measurement Model, [15] suggests that construct-irrelevant variance can be assessed by examining both dimensionality and fit of the measurement while significant gaps between the subsequent items provide indication of construct under-representation

III. Findings and Discussion

PCA performed on the residuals resulted in the second factor extracted had a strength of about 2 items. Therefore, it can be concluded that the second factor thus did not contain enough information that can pose a threat to the main construct. The Rasch analysis as presented in Table 1 found both means of infit MNSQ and outfit MNSQ were close to the expected value of 1.00. Inspection with individual items showed that infit MNSQ values ranged from 0.81 to 1.19 while outfit MNSQ values ranged from 0.76 to 1.30. The results supported the following: (1) the unidimensionality assumption of the construct was met, and (2) the scores demonstrated little variation from model expectation – that there was evidence of consistency between 588 examinees' response and 50 items on the scale and the model's expectations.

Reliability of item difficulty measures were high (.99) suggesting that the ordering of item difficulty was replicable

with other comparable sample of examinee. Meanwhile, consistency of examinees' measures (equivalent to Cronbach's alpha) was also high (.90), indicating that it was highly likely that the ordering of examinees proficiency can be replicated since most of the variance was attributed to true variance of the Mathematics proficiency construct. From the findings, threat regarding construct irrelevant-variance was minimum based on the dimensionality test as well as the within-range fit indices. In addition, based on Figure 1 (*the Wright Map*), no gaps of .5 *logits* or more [16] between subsequent items on the measured proficiency scale was reported. It showed that the MPT was broad enough to include important sub-dimensions of Mathematics proficiency construct.

Its limitation notwithstanding, the present study extends the understanding of how Rasch Model framework can be used in test development to measure a certain construct. In terms of construct definition and validity inquiry, traditionally, assessment merely involves generalization from performance on a sample of tasks to expected performance on the universe of tasks from which the sample was drawn. As provided in the present study, the use of Rasch Model offers opportunity to deal with core measurement issues such as construct validity as well as providing richer interpretation regarding examinee performance. Theoretically, this study has added more evidence in favor of the Rasch Model as having the capacity to resolve some of the rudimentary issues in measurement. However, in order for construct validity to hold, the model requires more evidence especially the corresponding between theoretical perspective and the observable behaviors. Test developers would have to have a thorough understanding of the measured construct especially information on relative difficulties of the items so that they can conceptualize the measured construct.

TABLE I: ITEM STATISTICS

Item Label	Measure (*logits*)	Score	Count	SE	Infit MNSQ	Outfit MNSQ
Q1	-1.13	558	375	0.10	0.95	0.87
Q2	-0.82	566	351	0.10	0.90	0.87
Q3	-1.41	543	388	0.10	0.86	0.74
Q4	1.77	370	97	0.13	1.00	0.96
Q5	2.00	306	78	0.14	0.95	0.94
Q6	0.90	493	173	0.11	1.09	1.15
Q7	-0.26	575	297	0.09	0.98	0.94
Q8	0.67	537	199	0.10	1.15	1.23
Q9	-0.19	572	288	0.09	0.93	0.90
Q10	0.10	576	261	0.09	1.19	1.24
Q11	-0.70	568	339	0.10	1.09	1.19
Q12	-0.21	570	289	0.09	0.99	1.02
Q13	-0.98	560	361	0.10	0.89	0.82
Q14	1.44	382	118	0.12	0.97	1.01
Q15	-0.59	572	331	0.10	0.92	0.92
Q16	-0.61	573	334	0.10	0.85	0.79
Q17	-0.60	572	333	0.10	1.05	1.10
Q18	0.02	573	269	0.09	1.16	1.28
Q19	0.63	538	204	0.10	1.03	1.09
Q20	1.09	468	154	0.11	1.19	1.30
Q21	-0.80	566	349	0.10	0.81	0.72
Q22	-0.20	576	292	0.09	0.88	0.87
Q23	0.07	575	266	0.10	0.86	0.84
Q24	-0.64	571	337	0.10	0.84	0.77
Q25	-0.17	575	288	0.09	1.11	1.13
Q26	-0.52	574	325	0.09	1.02	1.01
Q27	-0.27	576	299	0.09	0.93	0.96
Q28	-1.54	531	388	0.11	0.89	0.79
Q29	-1.65	518	386	0.11	0.90	0.82
Q30	0.26	568	243	0.10	1.09	1.10
Q31	0.56	537	210	0.10	1.03	1.07
Q32	-0.09	573	278	0.09	0.95	0.95
Q33	-0.53	574	326	0.09	0.97	0.94
Q34	-0.08	569	274	0.09	1.03	1.02
Q35	-0.38	573	309	0.09	1.01	1.06
Q36	0.08	576	264	0.10	1.06	1.09
Q37	1.37	426	127	0.12	0.98	1.08
Q38	1.22	467	142	0.11	0.97	1.09
Q39	0.02	571	268	0.10	0.98	0.96
Q40	-0.05	571	272	0.09	0.97	0.98
Q41	0.29	566	240	0.10	1.13	1.16
Q42	0.33	553	233	0.10	1.06	1.12
Q43	0.30	565	236	0.10	1.02	1.02
Q44	-0.92	558	354	0.10	0.93	0.91
Q45	0.23	569	247	0.10	1.15	1.28
Q46	-0.32	573	303	0.09	1.13	1.18
Q47	-0.86	564	354	0.10	1.07	1.21
Q48	1.71	400	102	0.13	1.14	1.29
Q49	-0.71	564	337	0.10	0.81	0.76
Q50	2.18	296	68	0.15	0.99	1.00

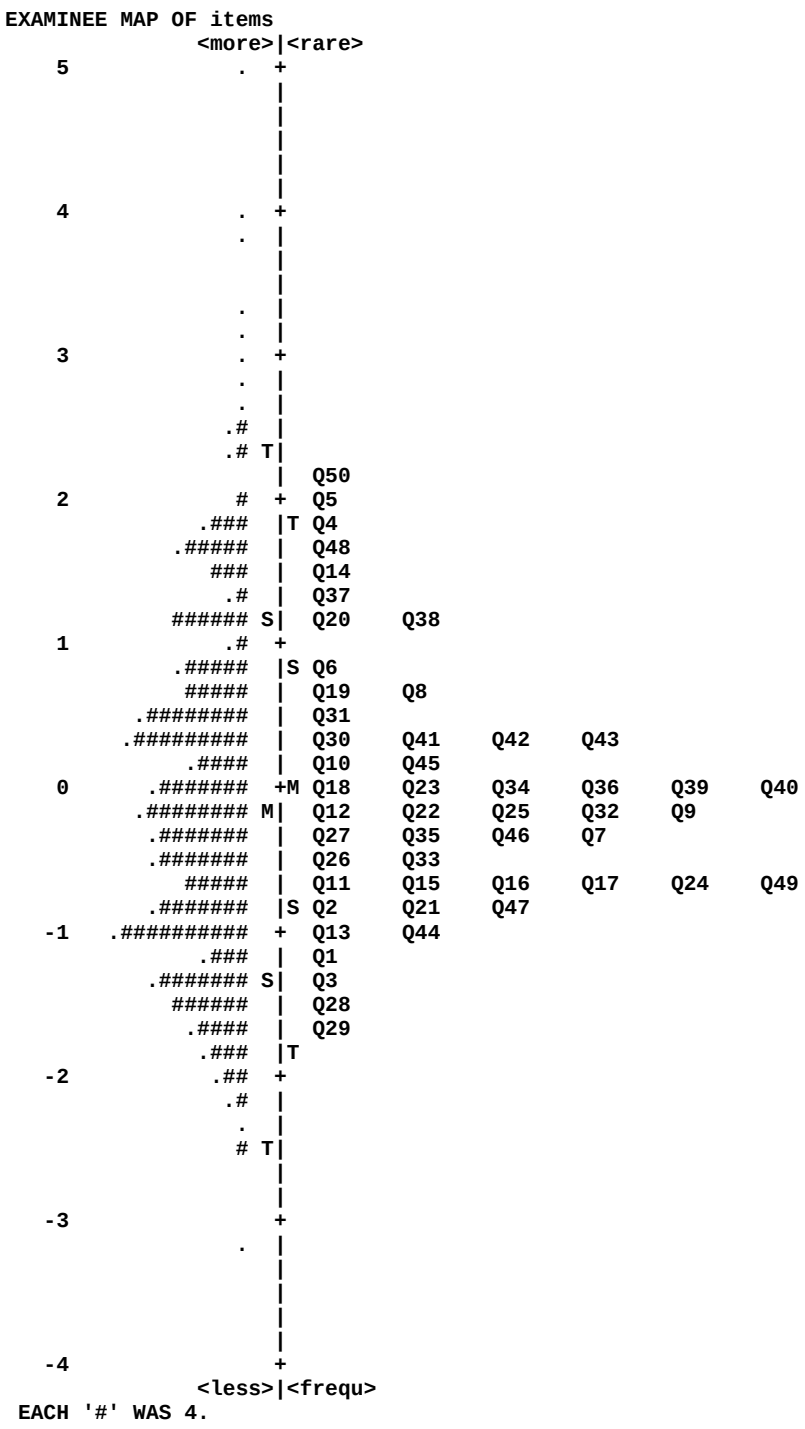

Fig.1. The wright map.

REFERENCES

[1] A. J. Nitko, *Educational Assessment of Students, 2nd. ed.* Englewood Cliffs, NJ: Merrill, 1996, ch. 1.

[2] L. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory.* Orlando, FL: Holt, Rinehart and Winston Inc, 1986, ch. 1.

[3] S. E. Embretson and S. P. Reise, *Item response theory for psychologists*, Mahwah, NJ: Lawrence-Erlbaum, 2000, ch. 1.

[4] S. M. Downing, "Item response theory: applications of modern test theory ", *Medical Education*, vol. 37, pp. 739-745, 2003.

[5] R. K. Hambleton and H. Swaminathan, *Item Response Theory: Principles and Applications*, Boston: Kluwer.Nijhoff, 1985, chap. 3.

[6] B. D. Wright and N. A. Panchapakesan, "A procedure for sample-free item analysis" in *Educational and Psychological Measurement*, vol.29, pp. 23-48, 1969.

[7] M. L. Stocking, Item response theory i*n Advances in Measurement in Educational Research and Assessment*, J. P. Keeves. And Masters, G. N. (eds.) Armsterdam : Pergamon, 1999, pp. 55-63

[8] B. D. Wright, "IRT in the 1990s: Which models work best?" in *Rasch Measurement Transaction*, vol 6: 1, pp. 1145-1146.

[9] J. P. Keeves, and S. Alagumalai, New approach to measurement in J. P. Keeves and G. N. Masters, G. N. (eds.) Armsterdam : Pergamon, 1999, pp. 23-42.

[10] D. Andrich, "The application of an unfolding model of the PIRT type to measurement of of attitude', *Applied Psychological Measurement*, vol. 12, pp. 33-5, 1992.

[11] T. G. Bond, and C. M. Fox, Applying *the Rasch model: Fundamental Measurement in Human Sciences, 1st ed*, Mahwah, NJ: Lawrence Erlbaum, 2001, chap. 2.

[12] Ministry of Education, *Curriculum Specifications for Mathematics Form 2* – 2002.

[13] J. M. Linacre, A user's guide to Winsteps – 2005

[14] S. Messick, Validity in R. L. Linn, (ed) *Educational Measurement* (3rd ed), Phoenix: Oryx Press, 1993, pp. 13-104.

[15] P. Baghaei, "Rasch Model as a construct validation tool" in. *Rasch Measurement Transaction*, vol 22: 1, pp. 1145-1146, 2008.

[16] J. M. Linacre, "When does a gap between measures matter?" in *Rasch Measurement Transaction*, vol 18: 3, p. 993, 2004.