Applying IRT Model in Validating a Dichotomously Scored Test Item

Arlene Nisperos Mendoza, Pangasinan State University, Philippines

The Asian Conference on Education 2022
Official Conference Proceedings

Abstract
This study conducted an item analysis to validate a dichotomously scored test using the Rasch measurement model, an Item Response Theory approach for test validation. It aimed to improve the quality of the test items in a departmentalized mathematics examination that had undergone content validity through subject experts. The response data gathered from randomly selected college students who had undergone the examination were fitted to the model. Rasch analysis revealed that the test appeared to be relatively difficult, indicating that the it needs to be revised further or that a better teaching strategy is needed to facilitate learning. It was also evident from the results that several misfit items appeared, and evidence of multidimensionality existed, which suggested that these items should be further modified, discarded or amended. However, both the item and person reliabilities were high. These findings suggest that an objective measurement for test validation, such as the Rasch measurement model, could help achieve greater precision in diagnosing test items and, consequently, construct a better measure for the assessment of students' abilities.

Keywords: Rasch Model, Item Response Theory, Item Analysis, Departmentalized Examination, Unidimensionality

Introduction

The goal of using tests in the teaching and learning process is accomplished only if the test is of good quality. Thus, developing valid and reliable tests is critical for evaluating student performance. The quality of any test was determined by analyzing the responses of students to any examination (Elee, L. I.; Onah, F. E. & Abanobi, C. C.,2018).

Classical Test Theory (CTT) and Item Response Theory (IRT) are two generally acceptable frameworks for evaluating the quality of tests in educational and psychological measurements (IRT). For approximately 100 years, classical test theory (CTT) has been extensively serving the testing field. Although CTT can provide important evidence for measuring instrument accuracy, several new psychometric tools may supplement or even replace this approach in collecting more accurate evidence to support inferences about the meaning and interpretation of scores. (Muñiz, 2017; by Zanon, et al., 2016).

The implementation of item response theory (IRT) in psychological and educational assessments has caused major and positive changes in psychological test development (Hambleton & Jodoin, 2003; Zanon et al., 2016). It has become a popular methodological framework for modeling response data from assessments in education and health; however, its use is not widespread among psychologists. (Zanon, et al., 2016).

IRT is a statistical theory composed of a variety of mathematical models that have the following characteristics: a) to predict person scores based on abilities or latent traits, and b) to establish a relationship between a person's item performance and the set of traits underlying item performance through a function called the "item characteristic curve" (Hambleton et al. 1991, Zanon, et al., 2016). The development of the Item Response Theory (IRT) addressed some weaknesses of CTT, namely: (1) the estimation of the test taker's ability does not depend on the characteristics of the tests used; (2) the item parameter estimation does not depend on the ability of the test taker; and (3) the measurement error can be searched for each individual (Susongko, 2016).

Item Response Theory models attempt to describe respondents' behavior based on their responses to each item. The logistic function is used to estimate the model in general, with three different formulations: 1PL (One Parameter Logistic model), 2PL (Two Parameter Logistic model), and 3PL (Three Parameter Logistic model). The 1PL model, also known as the Rasch Model after the Danish mathematician Georg Rasch, was used as a tool for item analysis in this study.

Rasch analysis is a psychometric technique developed to improve the precision with which researchers construct instruments, monitor instrument quality, and compute respondents' performance (Boone, 2016). The Rasch measurement model predicts how each person (the test-taker) should respond to each question based on the response data from the test's questions (referred to as items in this paper). In this analysis, both the test questions and the test takers are incorporated into a predictive mathematical model. That is, the difficulty level of the items and the ability level of the individuals are placed on a common scale so that the items and individuals can be easily compared (Karlin & Karlin, 2018).

Previous research has shown that analysis using the Rasch model is considered an appropriate and effective measurement technique for representing both students' ability to understand the material and the quality of the questions created (Runnels, 2012; Claesgens, et al., 2013;

Johnson, 2013; Boone, 2016; Talib, et al., 2018). One of the most fundamental ideas for understanding why Rasch theory is such an important tool for researchers is the concept of linearity (Boone, 2016). Based on their findings on the unexpected number of recommended modifications and deletions on the tests examined, Karlin and Karlin (2018) confirmed that the Rasch measurement model can be of tremendous value by offering greater precision in student assessment as well as greater assistance in test validation.

Rasch analysis is already being used by researchers in life sciences education to validate tests (Boone, 2016). Moreover, numerous global and local statistical tests have been proposed over the years to assess data conformity to the Rasch model principles (Baghaei, et al., 2017). However, most teachers are still unfamiliar with the approach's applicability to test improvement. As a result, the researcher finds it advantageous to employ this technique to explore its pertinence and improve measures in the assessment of their students.

Thus, the researcher attempted to do Rasch analysis for this purpose because this method maximizes the homogeneity of the trait and allows greater reduction of redundancy while sacrificing no measurement information by decreasing items and/or scoring levels to yield a more valid and simple measure (Bond & Fox, 2012). According to Bond and Fox (2001), the trait levels (the probability of a correct response or the probability of endorsing any option on each item) are modeled as a mathematical function of the difference between the person and the item parameters (Prieto-Adanes & Dias-Velasco, 2003; Zamora, et al., 2018).

The departmentalized examination in mathematics was used as an instrument in this study. This test was developed by mathematics and education experts and tested for content validity. It was administered to students who took Mathematics in the Modern World (MMW), which is one of the general education (GE) subjects mandated by the Commission on Higher Education (CHED Memorandum Order No. 20. S 2013) for college students who have completed the K to 12 programs. Because the test was multiple-choice, Rasch analysis on a dichotomously scored test was used. Its specific goals were to determine the item difficulty of each test item used in the test, diagnose the fit of the test items in the Rasch Model, assess the items' difficulty level against the students' level of ability, determine the item and person reliability, and evaluate the test's unidimensionality.

Participants

This study utilized the individual test results of 300 randomly chosen college students who took a Mathematics Achievement Test as a measure of their academic performance in Mathematics in the Modern World (MMW), a general education (GE) subject required by the Commission on Higher Education (CHED Memorandum Order No. 20. S 2013). The sample size was adequate to meet the requirements of the Rasch analysis (Linacre, 1994; Bond & Fox, 2012; Souza, 2017).

Procedure

After the Achievement Test was administered to all students, permission was requested from the head of the institution to secure copies of the test results. The student’s performance on each item was summarized, fitted to the Rasch model, and examined. A number right scoring method was applied in this study because the test is a multiple-choice type. That is, correct answers were scored with a positive value (1), and incorrect answers and absent or omitted answers with a value of zero (0). Hence the test was dichotomously scored.

In Rasch analysis, the probability of correctly answering an item can be expressed mathematically as the general statement (Bond & Fox, 2012):

$$P_{ni}(x = 1) = f(B_n - D_i) \quad (1)$$

Where P_n is the probability, x is any given score, and 1 is a correct response. This equation states that the probability (P_n) of a person n getting a score (x) of 1 on a given item (i) is a function (f) of the difference between a person's ability (B_n) and an item's difficulty (D_i). Given B_n and D_i, we can express then mathematically the function (f) expressing the probability of a successful response consisting of a natural logarithmic transformation of the person (B_n) and item (D_i) estimates as follows (Bond & Fox, 2012; Chan, et al., 2014; Sumintono, 2018; Winarti, et al., 2019):

$$P_{ni}\left(x_{ni} = 1/B_n, D_i\right) = \frac{e^{(B_n - D_i)}}{1 + e^{(B_n - D_i)}} \quad (2)$$

Where is the probability of person n on item i scoring a correct (x =1) response rather than an incorrect (x = 0) one, given person ability (B_n) and item difficulty (D_i). The given equation is the technical aspect of the Rasch model for dichotomously scored instruments. In this study, the Winsteps software was utilized, to facilitate the computation of the items' difficulty level, its fit to the Rasch model, the relationship between the items' level of difficulty and students' level of ability, and unidimensionality.

Data Collection

To gather quantitative data on student performance, a Mathematics Achievement Test constructed by a committee of faculty members specializing in Mathematics and Education was utilized in this study. This 50-item multiple-choice test covers the following topics: a) mathematics in our world (11 items), b) mathematical language and symbols (12 items), c) problem-solving and reasoning (5 items), and d) data management (22 items). This test was used to measure students' level of knowledge in Mathematics in the Modern World (MMW). MMW is a General Education (GE) subject taught to college students who have completed the added two years of high school under the K to 12 (Kindergarten to Grade 12) program (CHED Memorandum Order No. 20. S 2013). The instrument underwent content validity testing before test administration and item analysis. The test was subjected to the following process: (1) development of a Table of Specifications (TOS), (2) generation of an item pool, (3) review of the initial item pool by experts, (4) test administration, and (6) item analysis. The examinee was given one point for every correct choice of the letter that corresponded to the correct answer; hence, a total of 50 points were expected from each examinee.

Data Analysis

This study conducted an item analysis by fitting the individual raw scores to the Rasch model and the item statistics computed were analyzed.

The test items' level of difficulty estimates was expressed in logits, in which a logit value of 0 is arbitrarily set as the average or mean of item difficulty estimates (Bond and Fox, 2012). For many analyses, item difficulties range from −3 logits to +3 logits (Boone, 2016).

To diagnose the fit of the response data to the model, the infit and outfit statistics of each test item were examined. The fit statistics indicate where the test developer should decide whether to delete, restore, or reword an item. The value of the item's outfit and infit mean square and t statistics will be interpreted using the following range of the chi-square fit statistics (Wright & Linacre, 2002; Schumacker, 2004; Bond & Fox, 2012; Ee, et al., 2018; Kantahan, et al., 2020):

Mean Squares	tz	Response Pattern	Variation	Interpretation	Misfit Type
> 1.3	> 2.0	Too haphazard	Too much	Unpredictable	Underfit
< 0.75	< -2.0	Too determined	Too little	Guttman	Overfit

Table 1: Fit Statistics and Their General Interpretation

Moreover, an item characteristic curve (ICC) of the items was also presented to visualize the actual performances of the students on the items that overfit, underfit, and have a good fit to the model.

To assess the relations between the test item's level of difficulty and the student's level of ability, a plot of items according to their order of difficulty was examined through the item-person Wright map and its estimates computed.

The item and person reliability and separation indices were estimated from the simulation performed in the Rasch analysis. An item or person separation index of 1.5 (Ee, et al., 2018; Kantahan, 2018) and a reliability value higher than 0.70 was considered acceptable (Taber, 2018).

Unidimensionality was examined through Principal Component Analysis (PCA) of the residuals (Souza et al., 2017; Ee et al., 2018). PCA is one of the diagnoses by the Rasch model to ensure that all items share the same dimension, which is capable of sensing the ability of the instrument to measure the uniformity of single dimensions with acceptable noise levels (Linacre, 2012; Mokshein et al., 2019). For unidimensional measures, it is expected that the observed variance explained by the measures roughly matches the expected variance in the model. In addition, PCA analyses the components in the correlation matrix of the residuals (called contrasts). The "first contrast" is the component that explains the largest possible amount of variance in the residuals. The decision to consider a measure unidimensional or multidimensional is usually made by the researcher according to the purpose of the test. However, unexplained variances in the first contrast that are greater than 2.0 eigenvalue may indicate the presence of a second dimension (Souza et al., 2017; Ee et al., 2018).

Results and Discussion

This section summarizes the findings from a test assessment using the Rasch measurement model, including test item difficulty, test item fit to the model, test item difficulty and student ability relation, test item and person reliability, and unidimensionality.

Item Difficulty of the Individual Test Item

The item statistics from a Rasch analysis of a dichotomous test used in the study are shown in Table 2. It displays an ordered list of all the items (first column) based on their item difficulty measure (third column) and the associated logit error estimate (fourth column). The data in the second column showed the number of students who correctly answered the question.

Item	Raw Score	Difficulty Measure	Model S.E.	INFIT		OUTFIT		MISFIT TYPE
				MNSQ	ZSTD	MNSQ	ZSTD	
13	121	.01	.12	1.24	5.5	1.32	6.0	*Underfit*
14	50	1.35	.16	1.10	1.0	1.19	1.4	*Fit*
8	84	.62	.14	1.16	2.4	1.18	2.2	*Underfit*
28	145	-.35	.12	1.15	4.2	1.18	4.0	*Underfit*
5	87	.56	.13	1.13	2.1	1.14	1.8	*Underfit*
10	111	.16	.13	1.09	1.9	1.14	2.5	*Fit*
42	98	.37	.13	1.10	1.8	1.13	1.9	*Fit*
38	98	.37	.13	1.07	1.3	1.10	1.5	*Fit*
31	89	.53	.13	1.00	.0	1.08	1.1	*Fit*
22	81	.67	.14	1.07	1.1	1.07	.8	*Fit*
44	63	1.04	.15	1.00	.1	1.06	.6	*Fit*
48	147	-.38	.12	1.05	1.5	1.05	1.2	*Fit*
41	100	.34	.13	1.01	.1	1.05	.7	*Fit*
12	138	-.25	.12	1.04	1.2	1.04	1.0	*Fit*
23	237	1.87	.15	1.00	.1	1.04	.4	*Fit*
34	138	-.25	.12	1.02	.6	1.04	.9	*Fit*
33	216	1.46	.13	1.03	.6	1.01	.2	*Fit*
27	59	1.13	.15	1.03	.4	.99	.0	*Fit*
37	206	1.29	.13	1.00	.0	1.03	.4	*Fit*
6	66	.97	.15	.96	-.5	1.03	.3	*Fit*
32	157	-.53	.12	1.01	.4	1.03	.6	*Fit*
4	105	.26	.13	1.01	.3	.98	-.3	*Fit*
49	88	.55	.13	.99	-.2	1.01	.1	*Fit*
35	103	.29	.13	1.00	.1	.99	-.1	*Fit*
18	91	.49	.13	1.00	.0	1.00	.0	*Fit*
50	120	.02	.12	1.00	.0	.99	-.2	*Fit*
15	126	-.07	.12	1.00	-.1	.99	-.2	*Fit*
20	107	.23	.13	.99	-.2	1.00	.0	*Fit*
39	200	1.19	.13	1.00	-.1	.99	-.2	*Fit*
45	121	.01	.12	1.00	-.1	.99	-.1	*Fit*
47	195	1.11	.13	.97	-.6	.99	-.1	*Fit*
11	106	.24	.13	.99	-.3	.97	-.5	*Fit*
17	130	-.13	.12	.98	-.4	.98	-.4	*Fit*
25	153	-.47	.12	.98	-.5	.97	-.7	*Fit*
40	211	1.37	.13	.95	-1.0	.98	-.2	*Fit*
46	131	-.15	.12	.97	-.8	.96	1.0	*Fit*
43	82	.65	.14	.96	-.6	.96	-.4	*Fit*
19	169	-.71	.12	.95	-1.3	.94	1.2	*Fit*
3	154	-.49	.12	.95	-1.5	.93	1.6	*Fit*
29	80	.69	.14	.93	-1.0	.92	-.9	*Fit*
24	128	-.10	.12	.93	-1.7	.91	2.0	*Fit*
16	97	.39	.13	.93	-1.4	.93	1.1	*Fit*
30	112	.15	.13	.93	-1.6	.91	1.8	*Fit*
26	74	.81	.14	.91	-1.3	.92	-.9	*Fit*
36	196	1.12	.13	.90	-2.4	.91	1.4	*Overfit*
2	165	-.65	.12	.91	-2.9	.88	2.7	*Overfit*
1	185	-.95	.12	.90	-2.7	.85	2.7	*Overfit*
21	46	1.45	.17	.88	-1.1	.88	-.9	*Fit*
7	120	.02	.12	.87	-3.4	.85	3.2	*Overfit*
9	88	.55	.13	.86	-2.4	.84	2.3	*Overfit*
Mean	123.5	.00	.13	1.00	-.1	1.01	.1	
S.D.	46.1	.75	.01	.08	1.6	.10	1.6	

Table 2: Item Statistics: Misfit Order

Based on the value of the item difficulty estimates, which are expressed in logits, items 21, 14, 27, 44, and 6 appeared to be the most difficult items, having the highest value of the difficulty measure estimate. Item 21, which deals with the science to which the golden ratio belongs, is considered the most difficult item progressively, with the highest positive logit score.

Items 23, 33, 40, 37, and 39, on the other hand, are among the easiest because they have the lowest negative value of the item difficulty measure computed. Item 23, which covers inductive reasoning on number patterns and has negative logit estimates, was the most straightforward.

Figure 1 depicts the item characteristic curves (ICCs) for Item 23, the easiest item in this test, and Item 21, the most difficult item. It can be seen that the probability of success on Item 23 begins with a negative logit (-1.87), implying that even students with low ability have a chance of getting the item correct. In this case, there is a greater likelihood that more students will succeed in this item. Meanwhile, the ICC for item 21 shows that the likelihood of success on this item decreases from 1.5 to 7 logits. This only demonstrates that overcoming the difficulty of this item requires a high level of ability. As a result, students struggle to succeed in this item.

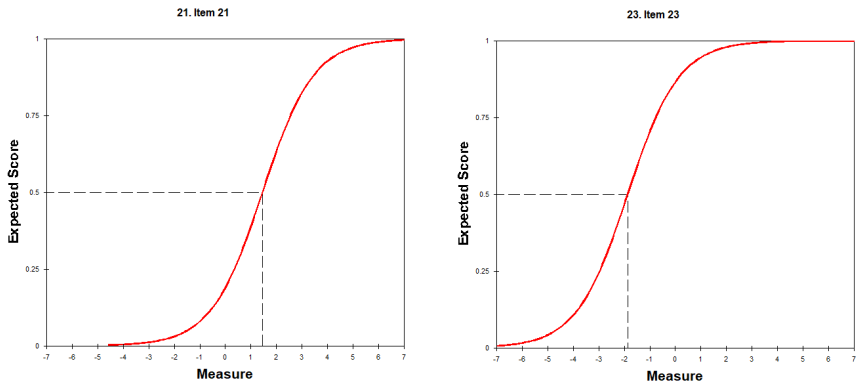

Figure 1: Item Characteristic Curves (or ICCs) for item 21 and item 23

In Rasch analysis, the mean of item difficulties was set to 0 points by default. Ignoring the measurement error for a moment, items 7, 13, 45, and 50 were calculated as having difficulty estimates that were closer to the exact value (0 logits). This means that the difficulty level of these items is within the range of the test-takers' abilities. In this case, the students have a 50 percent chance of correctly obtaining the item.

Test Item's Fit to the Rasch Model

The results of the individual test item fit statistics for both the unstandardized and standardized forms are also shown in Table 2. Mean squares are reported in the unstandardized form, while t-statistics are reported in the standardized form.

According to the findings, items 13, 28, 8, and 5 underfit the model because their infit t-values exceed 2.0 and their outfit t-values exceed 1.3. These positive fit statistics values indicate that the response string has more variation than expected; that is, it is less haphazard than expected. This means that a capable person gets easier items wrong unexpectedly, while

a less capable person gets harder items right unexpectedly. Similarly, the items' standardized mean square values revealed 13 to 24 percent more variation between the observed data and the model-predicted response pattern than would be expected if the data and the model were perfectly compatible. The presence of these underfit items in the test may degrade the test's quality; thus, the test developer should re-examine these items and scrutinize any mistakes or errors that may have been made in the test item's construction.

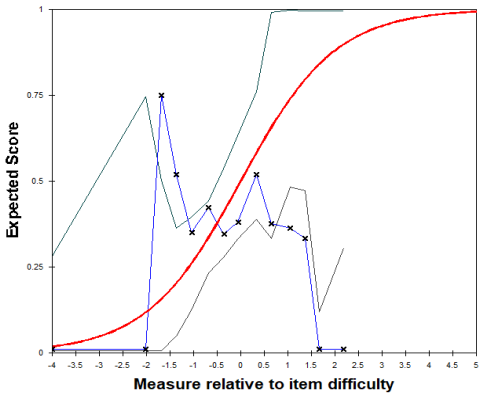

Figure 2. Actual person performance versus theoretical item characteristics curve (ICC) for underfit item 13

Figure 2 depicts the fit for the actual student performances (the jagged empirical ICC, blue curve) against the Rasch model expectations (the theoretical ICC, red curve) for item 13, the test item with too much variation and an almost exact difficulty estimate (0 logits).

The segment of the output for item 13 shows poor fit characteristics from the modeled expected, though its mean square value is not too bad. A poor fit indicates that the actual performance of the test-takers deviated from the modeled expectation. In this case, a slight deviation was observed between -2.0 and -1.0 logits and from a logits measure greater than 0.5. The curve also shows that students with a low level of ability have a higher chance of getting this item correct than those with a high level of ability, which is not expected.

Items 7, 2, 1, 9, and 36, on the other hand, overfit the model because their infit t values are less than -2 logits and their outfit statistic is less than 0.75 logits. These values indicate less variation than was modeled, implying that the response string is more similar to the Guttman-style response string, in which all easy items are correct and all difficult items are incorrect. The value of the infit mean square confirmed these findings, revealing a range of 9 to 14 percent less variation in the observed response pattern than was modeled. Although the presence of overfit items has few practical implications, test developers should be wary of their presence because these items may inflate test item reliability, leading us to conclude that the quality of our measure is higher than it is; additionally, the omission of overfitting items may rob the test of its best items; thus, a revision of these items should be performed first before recommending its deletion.

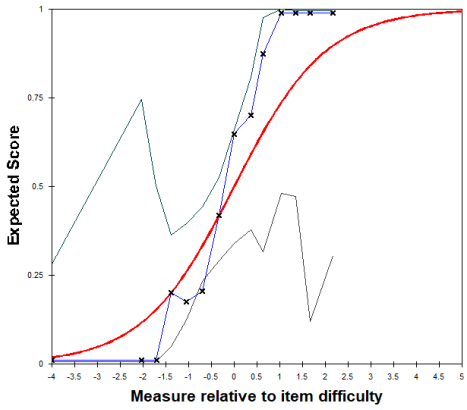

Figure 3. Actual person performance versus theoretical ICC for overfit item 7

Figure 3 displays the students' actual performance versus the theoretical item characteristics curve (ICC) for Item 7, a question with too little variation from the expected model. The figure shows that the students' actual performance ranges from 1.0 to 2.0 logits higher than those predicted. Similarly, their results below -1.5 logits deviated from the model, which was lower than expected. Overfitting for a low ability group indicates that the item can distinguish between minor differences in ability. An overfit is considered good in Classical Test Theory. According to Rasch theory, it is not a bad thing, but it usually indicates that something else is going on, such as item dependency (Linacre, 2017).

The item's difficulty level was at the midpoint of the test. As shown in Table 2, its mean square values are not too bad and are closer to the expected value of 1.0. The t-statistics, on the other hand, deviated farther from what was expected of the model. The item clearly follows the Guttman style in this case, as it overestimates the expectation based on the item's difficulty and the students' ability.

The findings on the underfit and overfit items only show that the observed data on these items do not conform to the Rasch Model because they have a value of outfit and infit t statistics that fall outside the acceptable range, despite having an almost exact difficulty level estimate. This simply means that these items are less compatible with the model than expected. Furthermore, these items imply the presence of multidimensionality; thus, they should be modified, discarded, or amended to focus on the target latent trait being tested. The evaluation of "fit" items to the Rasch Model ensured the measurement instrument's quality (Boone, Staver, and Yale, 2014).

Furthermore, while the mean of the unstandardized fit estimates (i.e., mean squares) computed is close to the expected value of 1, with the infit and outfit mean squares being close to that ideal, the mean and standard deviation of the standardized version of fit estimates (t statistics) show a slight deviation from the expected values of 0 and 1, respectively. These findings confirmed that the test was less compatible with the model's expectations due to these misfit items.

Items that satisfy the requirement for the mean square value and t statistics, on the other hand, are considered to have a good fit and be compatible with the modeled expected. Figure 4 depicts the students' actual performance versus the theoretical item characteristics curve (ICC) for item 50, one of the items with a good fit. The points plotted on the jagged curve (blue curve) represent the actual test performance of the 300 students. When the

performances fit perfectly to the Rasch model, the smooth curve (red curve) models the expected performance of the interaction between persons and the item (an impossible expectation).

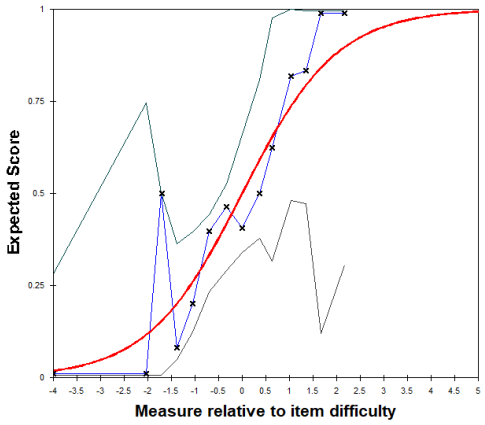

Figure 4: Actual person performance versus theoretical ICC for item 50

According to Figure 4, the students' performance in item 50, as reflected by the plotted points of their mean actual responses, is quite close to the Rasch model expectation of performance (the ICC). Although there was a slight deviation from the modeled curve around -2.0 logits relative to the item difficulty, this is the variation of actual around expected predicted by the Rasch model.

In this case, the infit and outfit mean square values for both items were close to 1.0, while their standardized versions, the infit and outfit t-statistics, were close to zero. These values only indicate the compatibility of the item with the model in addition to its difficulty measure estimate, which falls at the midpoint of the test with an almost exact value (0 logits).

Item Difficulty and Student Ability Relations

Figure 5 displays a wright map of the relationships between the students' ability and the difficulty of the items. As quality evidence, this graphical representation connects item difficulties and student ability estimates on a common scale; thus, both variables should match to justify that the test is maximally informative (Junpeng, 2020).

The distance of the step from the bottom of the path represents the difficulty of the item relative to other items. This is our representation of the item difficulty. Closer to the bottom is easier, and farther away is more difficult. Based on the representation of persons and items on the map, Item 33 was much more difficult than Item 23, whereas Item 21 was the most difficult on this test. In this test, most students did not succeed in Item 21. Item 23, however, was easy. In fact, the easiest on this test, and most of the students got Item 23, correct. These findings show that all items were useful for discriminating ability among students of this group since not everyone was successful on the easiest item or got the most difficult item wrong.

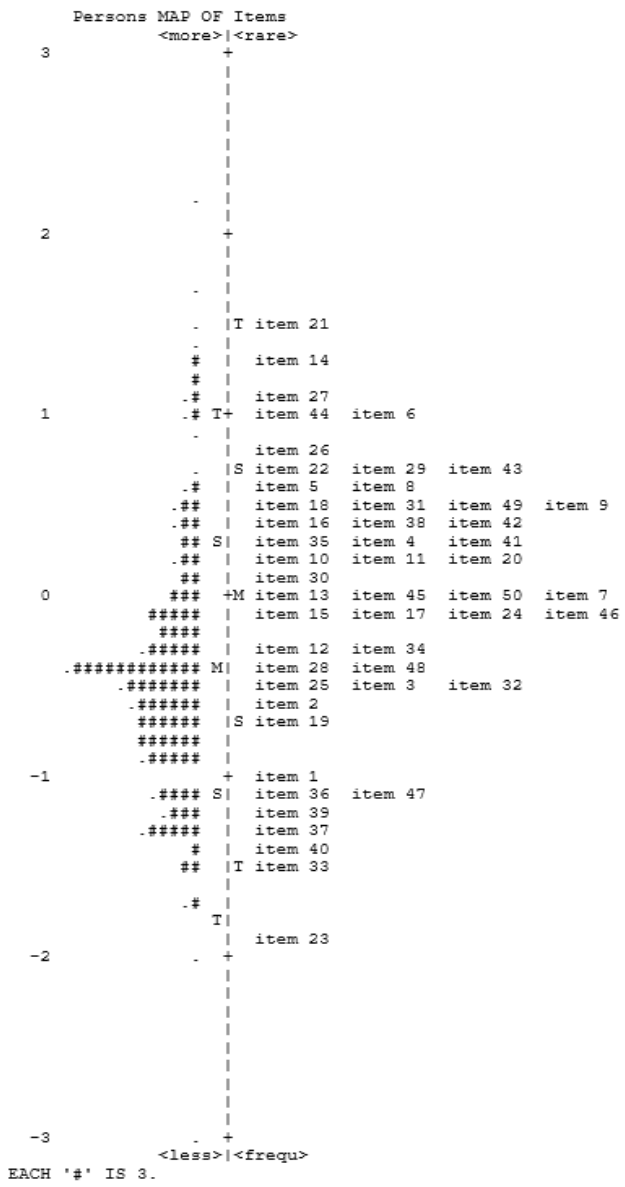

Figure 5: Item-person Wright Map

In addition, items 7, 13, 45, and 50 are located at 0 points in the item-person map for having an almost exact difficulty estimate (0 logits). Nine students had a 50 percent chance of getting these items correct. Furthermore, 57 more students could probably get these items right with more than a 50 percent probability of success. However, the remaining 78 percent of the total number of takers failed in these items.

This result shows that this type of test is somewhat a little bit difficult for the examinees' level of ability, although there are two students with a perfect score of 50. It can be observed that only one-third of the total samples have an equal to or higher than 50 percent chance of obtaining a correct answer on half of the total items. This means that the level of ability of the majority of examinees did not exceed the level of difficulty of the majority of the items. The figure shows that the majority of examinees failed the exam. This result could be regarded as a serious inadequacy in a test from a general test development perspective. The test requires even easier questions to raise the "ceiling" of the test. Otherwise, teachers need a better teaching strategy to facilitate their students' learning (Bond & Fox, 2012).

Item and Person Reliability

Table 3 presents the reliability of the test items. It includes an overall summary of the items' mean difficulty estimates, item and person reliability, and separation. The table shows a large positive value for the item difficulty mean estimate, which indicates that the test is difficult for the sample group of students who took the examination. This corroborates the results reflected in the item-person Wright map shown in Figure 5. The standard deviation of 46.1 for item estimates indicates a greater spread of item measures or variation in those measures than with person measures.

In this test, the reliability of the item difficulty estimates (0.97) was very high on a scale of 0 to 1, and was more than acceptable. Moreover, the item separation value of 5.54 expresses that the persons have differentiated more than five levels of item difficulty. These findings indicate that we can quite readily rely on this order of item estimates to be replicated when we administer the same test to other groups of students for whom it is suitable. According to Nielsen (2018), good measurements should have a high degree of reliability if the scores are consistent. However, the findings of the overfitting items may affect the level of reliability. Thus, further examination of the effects of these items on test reliability should be performed.

However, the person reliability index (0.77) is relatively high. This suggests that if the same group of persons were to be given another set of items measuring the same construct, almost the same estimate of a person's ability would be expected. The person separation of 1.81 states that items were able to differentiate between more than one level of a person's ability.

Statistics	Score (Item)	Score (Person)
Mean	123.5	20.6
S.D.	46.1	6.9
SD (adjusted)	0.74	0.60
Real RMSE	0.13	0.33
Item/Person Reliability	0.97	0.77
Item/Person Separation	5.54	1.81

Table 3. Summary of Item and Person Estimates

Unidimensionality of the Test

The unidimensionality of the test was examined through Principal Component of Analysis (PCA) of residuals in Rasch. Unidimensional means that the test only measures one's ability (Susongko, 2016). Tables 4 and 5 summarize these findings.

The data in Table 4 reveal that there was a total variance of 65.9 eigenvalue units in the observations. Of this total variance, 15.9 eigenvalue units were explained by person and item measures. Meanwhile, the unexplained variance had 50 eigenvalue units that covered more than 75 percent of the total variance. This value varies from the result of the Rasch measure and the difference is significant. This obtained value of the unexplained variance from other sources could be anything not meant to be included in the test; hence, it was not explained by the Rasch measurement.

			Empirical		Modeled
Total variance in observations	=	65.9	100.0%		100.0%
Variance explained by measures	=	15.9	24.2%		24.6%
Unexplained variance (total)	=	50.0	75.8%	100.0%	75.4%
Unexplained variance in 1st contrast	=	3.2	4.8%	6.4%	
Unexplained variance in 2nd contrast	=	2.5	3.8%	5.0%	
Unexplained variance in 3rd contrast	=	2.1	3.2%	4.2%	
Unexplained variance in 4th contrast	=	2.0	3.0%	4.0%	
Unexplained variance in 5th contrast	=	1.7	2.5%	3.3%	

Table 4. Table of Standardized Residual Variance (in Eigenvalue Units)

Based on the results, substantive structures contribute to unexplained variance. After extracting information from the data through Rasch measurement, residuals were left, as reflected in the contrasts indicated in the findings. Residuals are the observed performance of students on the item minus what is expected by the Rasch model. The smaller the residuals, the better the data fits the model (Kazemi, 2020). In this case, the results show five contrasts, and some consist of more than 2 eigenvalue units which indicates the existence of potential dimension (Linacre, 1998; Ee, Yeo, and Kosnin, 2018). Hence, a principal component analysis of Rasch residuals (PCAR) or linearized Rasch residuals was employed to extract meaningful information from these contrasts. Table 5 presents a summary of the findings for the first factor, which had the highest factor sensitivity ratio among the contrasts.

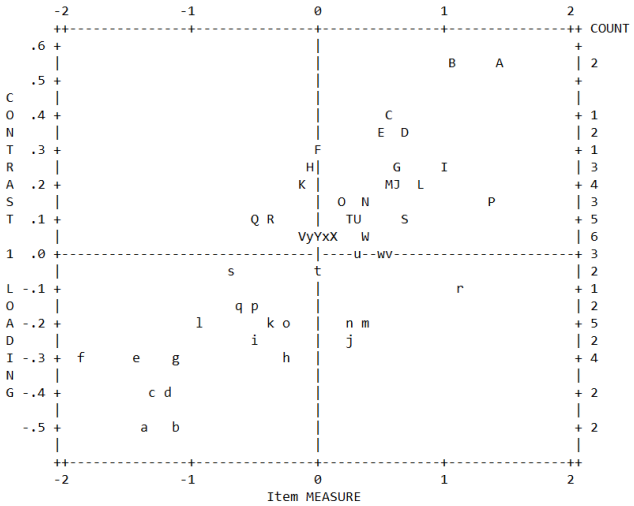

Figure 6. Standardized Residual Variance Scree Plot (Contrast 1)

In the first contrast, the unexplained variance was 3.2 units which means that there were around three eigenvalues creating a subdimension in the data. These items have something in common other than the Rasch dimension, which clusters these items together. Figure 6 shows the factor plot of the standardized residuals after the primary Rasch dimension was extracted. It can be seen that items 21(A), 44(B), and 9(C) have higher factor loadings and can be seen at the top of the map. These items have substantial variance that remains unexplained by the primary Rasch measure.

Table 5 lists the factor loadings for the first dimension (contrast 1). These loadings indicate three items (21, 44, and 9) with substantial positive loadings on the factor discovered in the item residuals (i.e., with an off-dimension loading of 0.4 or greater). On the other hand, two items (40 and 36) are negatively correlated with the factor.

Loading	Measure	Infit Mnsq	Outfit Mnsq	Entry Number	Item
.55	1.45	.88	.88	A 21	Item 21
.53	1.04	1.00	1.06	B 44	Item 44
.42	.55	.86	.84	C 9	Item 9
.37	.69	.93	.92	D 29	Item 29
.36	.49	1.00	1.00	E 18	Item 18
.28	.02	1.00	.99	F 50	Item 50
.27	.65	.96	.96	G 43	Item 43
.27	-.07	1.00	.99	H 15	Item 15
.27	.97	.96	1.03	I 6	Item 6
.21	.62	1.16	1.18	J 8	Item 8
.21	-.13	.98	.98	K 17	Item 17
.19	.81	.91	.92	L 26	Item 26
.19	.55	.99	1.01	M 49	Item 49
.17	.39	.93	.93	N 16	Item 16
.17	.16	1.09	1.14	O 10	Item 10
.15	1.35	1.10	1.19	P 14	Item 14
.11	-.47	.98	.97	Q 25	Item 25
.10	-.38	1.05	1.05	R 48	Item 48
.10	.67	1.07	1.07	S 22	Item 22
.10	.23	.99	1.00	T 20	Item 20
.08	.29	1.00	.99	U 35	Item 35
.07	-.10	.93	.91	V 24	Item 24
.06	.37	1.07	1.10	W 38	Item 38
.06	.15	.93	.91	X 30	Item 30
.04	.02	.87	.85	Y 7	Item 7
.04	-.15	.97	.96	y 46	Item 46
.03	.01	1.00	.99	x 45	Item 45
.00	.53	1.00	1.08	w 31	Item 31
-.50	-1.37	.95	.98	a 40	Item 40
-.50	-1.12	.90	.91	b 36	Item 36
-.39	-1.29	1.00	1.03	c 37	Item 37
-.38	-1.19	1.00	.99	d 39	Item 39
-.32	-1.46	1.03	1.01	e 33	Item 33
-.31	-1.87	1.00	1.04	f 23	Item 23
-.29	-1.11	.97	.99	g 47	Item 47
-.29	-.25	1.02	1.04	h 34	Item 34
-.27	-.53	1.01	1.03	i 32	Item 32
-.24	.24	.99	.97	j 11	Item 11
-.22	-.35	1.15	1.18	k 28	Item 28
-.21	-.95	.90	.85	l 1	Item 1
-.18	.37	1.10	1.13	m 42	Item 42
-.18	.26	1.01	.98	n 4	Item 4
-.18	-.25	1.04	1.04	o 12	Item 12
-.17	-.49	.95	.93	p 3	Item 3
-.14	-.65	.91	.88	q 2	Item 2
-.10	1.13	1.03	.99	r 27	Item 27
-.06	-.71	.95	.94	s 19	Item 19
-.06	.01	1.24	1.32	t 13	Item 13
-.01	.34	1.01	1.05	u 41	Item 41
.00	.56	1.13	1.14	v 5	Item 5

Table 5: Principal Component Analysis of Standardized Residual Correlations for Items on First Dimension (Sorted By Loading)

These findings provide empirical evidence for the existence of a separate subscale. However, it is up to the researcher to decide whether this is sufficiently large and meaningful to measure separately from Rasch measures. The costs/benefits of including these items as part of the original Rasch dimension, hence potentially losing some sensitivity or validity of the

measurement, or excluding these items from the total score and works towards assessing and interpreting the other dimension separately should be reflected.

Conclusion

This study focuses on the application of the Rasch model, an IRT approach for test item analysis. Based on the findings, the test appeared to be difficult based on the takers' level. Hence, there is a need to construct easier questions or better teaching strategies to facilitate the learning of their students.

In addition, further examination of the effect of the overfit items on the level of test reliability was suggested by the analysis. Based on the findings, these recommended modifications of the test show that even if a test had already undergone content validity through experts in the given field, the Rasch measurement model can be of tremendous value by offering greater precision in diagnosing and validating a test, as well as in the assessment of the students (Karlin & Karlin, 2018).

Furthermore, it shows that the examination subjected to Rasch analysis still had some misfit items. Moreover, several substantive structures contributed to the unexplained variance. These findings provide empirical evidence for the existence of a separate subscale or multidimensionality, which suggests a modification, discarding, or amendment of the misfit items, focusing on the target latent trait being tested.

This research demonstrates the importance of applying an Item Response Theory (IRT) approach to item analysis. In this case, Rasch analysis was applied to introduce teachers to one of the robust tests that can be used for item analysis. In addition, most of the test being constructed was of a multiple-choice type; hence, this study found it beneficial to let teachers explore and learn the IRT approach applicable to a dichotomously scored test.

References

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a descriptive fit statistic for the rasch model. *North American Journal of Psychology*, *19*(1), 155-168.

Bond, T., & Fox, C. (2012). Applying the rasch model, fundamental measurement in the human sciences. *Lawrence Erlbaum Associates, Inc*. (2nd Ed.).

Boone, W. (2016). Rasch analysis for instrument development: why, when, and how?. *CBE Life Sciences Education*, *15*(4), 1-7. https://doi.org/10.1187/cbe.16-04-0148

Boone, W., Staver, J., & Yale, M. (2014). *Rasch Analysis in the Human Sciences*. The Netherlands: Springer, 2014. xvi+482 pp. ISBN 978-94-007-6856-7.

Claesgens, J., Scalise, K., & Stacy, A. (2013). Mapping student understanding in chemistry: The perspectives of chemists. *Educación Química*, *24*(4), 407-415. https://doi.org/10.1016/S0187-893X(13)72494-7

Commission on Higher Education (CHED), (2013). General Education Curriculum: Holistic Understandings, Intellectuals and Civic Competencies. CHED Memorandum Order No. 20. S 2013.

Ee, N., Yeo, K., & Kosnin, A. (2018). Item analysis for the adapted motivation scale using rasch model. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(4), 264-269. http://doi.org/10.11591/ijere.v7i4.15376

Eleje, L. I.; Onah, F. E. & Abanobi, C. C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test ıtem analysis results. European Journal of Educational and Social Sciences, 3 (1), 57 -75.

Johnson, P. (2013). Una progresión de aprendizaje para la comprensión del cambio químico. *Educacion Quimica*, *24*(4), 365–372. http://dx.doi.org/10.1016/S0187-893X(13)72489-3

Junpeng, P., et al. (2020). Validation of a digital tool for diagnosing mathematical proficiency. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(3), 665-674. http://doi.org/10.11591/ijere.v9i3.20503

Karlin, O., & Karlin, S. (2018). Making better tests with the rasch measurement. *Insight: A Journal of Scholarly Teaching*, *13*, 76-100.https://doi.org/10.46504/14201805ka

Kazemi, S., et al. (2020). Development and validation of a null curriculum questionnaire focusing on 21st century skills using the Rasch model. *Cogent Education*, *7*(1). https://doi.org/10.1080/2331186X.2020.1736849

Linacre, J. (2017). *Winsteps Rasch measurement computer program*. Beaverton, OR: Winsteps.com.

Nielsen, T. (2018). The intrinsic and extrinsic motivation subscales of the motivated strategies for learning questionnaire: A Rasch-based construct validity study. *Cogent Education*, *5*(1), 1-19. https://doi.org/10.1080/2331186X.2018.1504485

Runnels, J. (2012). Using the rash model to validate a multiple choice English achievement test", *International Journal of Language Studies*, 6(4), 141-155.

Susongko, P. (2016). Validation of science achievement test with the rasch model. *Journal Pendidikan IPA Indonesia, 5*(2), 268-277. https://dx.doi.org/10.15294/jpii.v5i2.7690

Talib, A., Alomary, F., & Alwadi, H. (2018). Assessment of student performance for course examination using rasch measurement model: a case study of information technology fundamentals course. *Education Research International*, 1–8. https://doi.org/10.1155/2018/8719012

Zamora-Araya, J. A., Smith-Castro, V., Montero-Rojas, E., & Moreira-Mora, T. E. (2018). *Advantages of the Rasch Model for analysis and interpretation of attitudes: The case of the Benevolent Sexism Subscale*. Revista Evaluar, 18(3), 1-13. Available at https://revistas.unc.edu.ar/index.php/revaluar

Zanon, C., Hutz, C.S., Yoo, H.,. et al., (2016). *An application of item response theory to psychological test development. Psicologia*: Reflexão e Crítica volume 29, Article number: 18 (2016). Retrieved from https://doi.org/10.1186/s41155-016-0040-x

Contact email: arlenenmendoza1@gmail.com