Available online at https://www.guinnesspress.org

Recent Educational Research

Volume 2. Issue 1, PP:8-21

DOI: 10.59762/rer904105362120240305124511

ISSN: 2996-2366

Item Analysis Using Rasch Model: An Advancement For Test And Students' Ability Evaluation

Arlene Nisperos Mendoza[1*]

[1]Pangasinan State University, Philippines

ARTICLE INFO

Article History:

Received: Aug-16-2023

Revised: Dec-20-2023

Accepted: Jan-24-2024

Corresponding Author:

Arlene Nisperos Mendoza

Pangasinan State University, Philippines

arlenenmendoza1@gmail.com

ABSTRACT

This study used the Rasch measurement model to assess the quality of the items in a college-level mathematics achievement test. It aimed to improve these items further after the experts evaluated their content validity. Examining the item difficulty of the individual test items, the fit of the items to the model, the relation of the item difficulty to the student ability level, item and person reliability, and the unidimensionality of the test comprised the preliminary analysis. The findings revealed that the test was relatively difficult for the test-takers. Its item reliability was acceptable; however, the person reliability index suggested additional items. Moreover, misfit items and evidence of multidimensionality appeared. These findings indicate that the test should be revised to improve its reliability and validity. Furthermore, an objective measurement, such as the Rasch model, was recommended to achieve greater precision in diagnosing test items and, consequently, construct a better student measure.

Keywords: Rasch model, Item analysis, Item difficulty, Item reliability, Person reliability, Mathematics achievement, Unidimensionality.

INTRODUCTION:

In order to accurately measure teaching and learning processes and evaluate student performance, it is essential to construct a reliable test (Fatimah, et al. (2020). When constructing tests, it is important to carefully evaluate content validity to ensure that the test accurately assesses content knowledge. Additionally, an item analysis should be conducted to assess the quality of individual test items and the overall test. This involves using both statistics and expert judgment to enhance the quality of the test, result ing in valid and reliable results that accurately quantify students' abilities (Susongko, 2016).

When developing measures to evaluate students' performance, teachers typically aim to use raw scores or ratings to indicate the extent of their students' competence in the target area. To ensure the validity of the measure, it is important for teachers to confirm that the set of items included in the measure assesses only the intended trait, meaning it is unidimensional. In order to achieve this goal, the researcher conducted Rasch analysis, which maximizes the unidimensionality of the trait, and allows for greater reduction of redundancy without compromising measurement accuracy. This is achieved by decreasing the number

of items and/or scoring levels to produce a simpler and more valid measure (Bond & Fox, 2012). Rasch analysis is a psychometric method that enhances the precision of instrument construction, quality monitoring, and computation of respondents' performance (Boone, 2016). The Rasch measurement model utilizes response data from test questions (known as items) to predict how each test-taker should answer each question. This analysis involves incorporating both the test questions and the test-takers into a mathematical model that predicts their performance. The items' difficulty level and the test-takers' ability level are placed on a shared scale, allowing for a straightforward comparison of both items and test-takers (Karlin & Karlin, 2018).

Prior research has demonstrated that utilizing the Rasch model for analysis is a suitable and effective technique for measuring both students' ability to comprehend the material and the quality of test questions created (Runnels, 2012; Claesgens, et al., 2013; Johnson, 2013; Boone, 2016; Talib, et al., 2018). The concept of linearity is a fundamental principle that explains the significance of Rasch theory as a research tool (Boone, 2016). In a study conducted by Karlin & Karlin (2018), they confirmed the value of the Rasch measurement model by demonstrating its ability to offer more precise assessments of students and better validation of tests. This was evident from their findings, which identified an unexpected number of recommended modifications and deletions in the examined tests.

Life science education researchers have already employed Rasch analysis to validate tests (Boone, 2016). Over time, several global and local statistical tests have been proposed to verify data conformity with the principles of the Rasch model (Baghaei, et al., 2017). Despite this, many teachers have not yet embraced this approach for enhancing tests. Consequently, the researcher considers it advantageous to apply this technique to assess its suitability and enhance the measures used to evaluate their students. The purpose of this study was to utilize the Rasch Model to conduct item analysis on dichotomously scored items and refine the test items that experts had previously validated. The research utilized a Mathematics Achievement Test to evaluate students' performance in Mathematics in the Modern World, a general education subject required by the Commission on Higher Education (CHED) for college students who have completed the K to 12 program. As the test items were in a multiple-choice format, Rasch analysis of dichotomous items was employed. The study examined the item difficulty of individual test items included in the test, evaluated the fit of the test items in the Rasch Model, assessed the difficulty level of the items relative to the students' ability level, determined the item and person reliability, and evaluated the unidimensionality of the test. The goal was to enhance the accuracy and validity of the test items for improved student measurement.

Design and Methods

Research Design

The goal of this study was to assess the quality of the test items within a Mathematics Achievement test in terms of their reliability and item characteristics. To achieve this, a cross-sectional study was conducted utilizing Rasch analysis.

Participants

To assess academic success in Mathematics in the Modern World (MMW), a general education (GE) subject mandated by the Commission on Higher Education (CHED Memorandum Order No. 20. S 2013), this study utilized the individual test results of 300 randomly selected college students who had taken the Mathematics Achievement Test. The sample size was deemed sufficient to meet the criteria required for Rasch analysis, as outlined by Linacre (1994), Bond and Fox (2012), and Souza (2017).

Procedure

Upon administering the Achievement Test to all students, the researchers requested permission from the head of the institution to obtain a copy of the test results. Each student's performance on every item was summarized, aligned with the Rasch model, and analyzed. In this study, a number-right scoring method was utilized since the test consisted of multiple-choice questions. This method assigns a positive value (1) to correct responses and a value of zero (0) to incorrect and unanswered items. Conse-

quently, the test was dichotomously scored. The Rasch analysis involves expressing the likelihood of correctly answering an item in a mathematical format, which can be stated generally as per Bond and 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. According to Bond and Fox (2012), this equation signifies that the likelihood (P_n) of an individual (n) obtaining a score (x) of 1 on a specific item (i) is reliant on the disparity between the individual's ability (B_n) and the item's difficulty (D_i). By mathematically expressing the person (B_n) and item (D_i) estimates using a natural logarithmic transformation as given below, we can determine the probability of a successful response, as per Chan et al. (2014), Sumintono (2018), and 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 $P_{ni}\left(x_{ni} = 1/B_n, D_i\right)$ 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 provided equation is the technical component of the Rasch model for instruments that are dichotomously scored. To analyze the items' difficulty level, their alignment with the Rasch model, the relationship between item difficulty and students' ability level, and the unidimensionality in this study, the Winsteps software was utilized for computation.

Data Collection

In this study, a Mathematics Achievement Test was utilized to obtain quantitative data on the students' performance. The test consisted of 50 multiple-choice items covering the following topics: mathematics in our world (11 items), mathematical language and symbols (12 items), problem-solving and reasoning (5 items), and data management (22 items). This test was employed to measure the student's level of knowledge in Mathematics in the Modern World (MMW), which is one of the General Education (GE) subjects taught to college students who have completed the K to 12 (Kindergarten to Grade 12) program, according to the Commission on Higher Education's Memorandum Order No. 20. S 2013. Prior to test administration and item analysis, the instrument underwent content validity testing, which involved the development of a Table of Specifications (TOS), item pool generation, expert review of the initial item pool, test administration, and item analysis. Each correct answer was scored with one point, and a maximum of 50 points was expected from each examinee.

Data Analysis

The item analysis in this study involved. Fitting the individual raw scores to the Rasch model and analyzing the computed item statistics The difficulty level estimates of the test items were expressed in logits, where a value of 0 logits was set as the average or mean of the item difficulty estimates, as per Bond and Fox (2012). In most cases, item difficulties range from −3 logits to +3 logits, as stated by Boone (2016). In order to evaluate the conformity of the response data with the model, the infit and outfit statistics for each test item were analyzed. These fit statistics assist the test developer in deciding whether to revise, remove, or modify an item. The chi-square fit statistics range provided by Wright and Linacre (2002), Schumacker (2004), Bond and Fox (2012), Ee et al. (2018), and Kantahan et al. (2020) will be utilized to interpret the item's outfit and infit mean square and t statistics:

Additionally, an item characteristic curve (ICC) was provided to illustrate the students' actual performances on items that overfit, underfit, and fit well with the model. This allows for a clear visualization of the items' characteristics. In order to evaluate the relationship between the level of difficulty of the test items and the students' ability level, the item-person Wright map was utilized. This involved examining a plot of the items arranged in order of difficulty and computing its estimates. Using the simulation conducted in the Rasch analysis, the item and person reliability and separation indices were estimated. An acceptable value for the separation index for items or persons is 1.5, as per Ee et al. (2018) and Kantahan

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

(2018), while a reliability value of 0.70 or higher is considered acceptable, according to Taber (2018).

This study employed Principal Component Analysis (PCA) of the residuals to assess unidimensionality, following the recommendation of Souza et al. (2017) and Ee et al. (2018). PCA is a diagnostic tool used in the Rasch model to ensure that all items measure the same construct consistently. In unidimensional measures, the expected variance should match the observed variance, and the "first contrast" explains the largest amount of variance. Researchers decide whether a measure is unidimensional or multidimensional based on its purpose, but if the unexplained variances in the first contrast are greater than 2.0 eigenvalue, this may suggest the presence of a second dimension (Souza et al., 2017; Ee et al., 2018).

Results and Discussion

This section offers a condensed overview of the outcomes acquired from a test evaluation using the Rasch measurement model. The summary encompasses diverse facets, such as the level of difficulty of the test items, their adherence to the model, the interconnection between item difficulty and student proficiency, and the dependability of both the test items and the individuals subjected to testing. Furthermore, the section scrutinizes the degree to which the test is characterized by a unidimensional construct.

Item Difficulty of the Individual Test Item

The item statistics obtained from a Rasch analysis of a dichotomous test utilized in the study were displayed in Table 2. The table includes a comprehensive list of all the items (first column), along with their corresponding item difficulty measures (third column) and associated logit error estimates (fourth column). The data in the second column of the table represents the number of students who answered each question correctly.

According to the values of the item difficulty estimates, which are expressed in logits, items 21, 14, 27, 44, and 6 were determined to be the most challenging items, possessing the highest difficulty measure estimate. Among these items, item 21, which assesses knowledge of the science of the golden ratio, was found to be the most arduous, progressively increasing in difficulty with the highest positive logit score.In contrast, items 23, 33, 40, 37, and 39 were classified as the easiest as they exhibited the lowest negative value of the computed item difficulty measure. Notably, item 23, which evaluated inductive reasoning on number patterns and was determined to have negative logit estimates, was identified as the most straightforward item among them. Figure 1 portrays the item characteristic curves (ICCs) for Item 23, the easiest item in the test, and Item 21, the most challenging item. The ICCs reveal that the probability of success on Item 23 begins with a negative logit (-1.87), indicating that even students with low ability have a fair chance of answering the item correctly. As a result, there is a higher possibility that a greater number of students will successfully answer this item. Conversely, the ICC for Item 21 displays that the likelihood of success on this item decreases as the logit values increase from 1.5 to 7. This finding implies that a high level of ability is required to overcome the difficulty of this item. Consequently, students find it challenging to answer this item accurately.

In Rasch analysis, the default setting is to set the mean of item difficulties to 0 points. Ignoring the influence of measurement error, items 7, 13, 45, and 50 were identified as having

Item	Raw Score	Difficulty Measure	Model S.E.	INFIT		OUTFIT	
				MNSQ	ZSTD	MNSQ	ZSTD
21	46	1.45	0.17	0.88	-1.1	0.88	-0.9
14	50	1.35	0.16	1.1	1	1.19	1.4
27	59	1.13	0.15	1.03	0.4	0.99	0
44	63	1.04	0.15	1	0.1	1.06	0.6
6	66	0.97	0.15	0.96	-0.5	1.03	0.3
26	74	0.81	0.14	0.91	-1.3	0.92	-0.9
29	80	0.69	0.14	0.93	-1	0.92	-0.9
22	81	0.67	0.14	1.07	1.1	1.07	0.8
43	82	0.65	0.14	0.96	-0.6	0.96	-0.4
8	84	0.62	0.14	1.16	2.4	1.18	2.2
5	87	0.56	0.13	1.13	2.1	1.14	1.8
9	88	0.55	0.13	0.86	-2.4	0.84	-2.3
49	88	0.55	0.13	0.99	-0.2	1.01	0.1
31	89	0.53	0.13	1	0	1.08	1.1
18	91	0.49	0.13	1	0	1	0
16	97	0.39	0.13	0.93	-1.4	0.93	-1.1
38	98	0.37	0.13	1.07	1.3	1.1	1.5
42	98	0.37	0.13	1.1	1.8	1.13	1.9
41	100	0.34	0.13	1.01	0.1	1.05	0.7
35	103	0.29	0.13	1	0.1	0.99	-0.1
4	105	0.26	0.13	1.01	0.3	0.98	-0.3
11	106	0.24	0.13	0.99	-0.3	0.97	-0.5
20	107	0.23	0.13	0.99	-0.2	1	0
10	111	0.16	0.13	1.09	1.9	1.14	2.5
30	112	0.15	0.13	0.93	-1.6	0.91	-1.8
7	120	0.02	0.12	0.87	-3.4	0.85	-3.2
50	120	0.02	0.12	1	0	0.99	-0.2
13	121	0.01	0.12	1.24	5.5	1.32	6
45	121	0.01	0.12	1	-0.1	0.99	-0.1
15	126	-0.07	0.12	1	-0.1	0.99	-0.2
24	128	-0.1	0.12	0.93	-1.7	0.91	-2
17	130	-0.13	0.12	0.98	-0.4	0.98	-0.4
46	131	-0.15	0.12	0.97	-0.8	0.96	-1
12	138	-0.25	0.12	1.04	1.2	1.04	1
34	138	-0.25	0.12	1.02	0.6	1.04	0.9
28	145	-0.35	0.12	1.15	4.2	1.18	4
48	147	-0.38	0.12	1.05	1.5	1.05	1.2
25	153	-0.47	0.12	0.98	-0.5	0.97	-0.7
3	154	-0.49	0.12	0.95	-1.5	0.93	-1.6
32	157	-0.53	0.12	1.01	0.4	1.03	0.6
2	165	-0.65	0.12	0.91	-2.9	0.88	-2.7
19	169	-0.71	0.12	0.95	-1.3	0.94	-1.2
1	185	-0.95	0.12	0.9	-2.7	0.85	-2.7
47	195	-1.11	0.13	0.97	-0.6	0.99	-0.1
36	196	-1.12	0.13	0.9	-2.4	0.91	-1.4
39	200	-1.19	0.13	1	-0.1	0.99	-0.2
37	206	-1.29	0.13	1	0	1.03	0.4
40	211	-1.37	0.13	0.95	-1	0.98	-0.2
33	216	-1.46	0.13	1.03	0.6	1.01	0.2
23	237	-1.87	0.15	1	0.1	1.04	0.4
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: Measure Order

difficulty estimates that were more proximate to the exact value (0 logits). This suggests that the difficulty level of these items falls within the range of the abilities of the test-takers. Consequently, students have a 50 percent chance of correctly answering these items.

Test Item's Fit to the Rasch Model

The outcomes of the individual test item fit statistics for both the unstandardized and standardized forms, which have been presented and tabulated in Table 2. These results have been summarized in Table 3. The mean squares have been reported in the unstandardized form, whereas the t-statistics have been reported in the standardized form. Based on the findings, it can be concluded

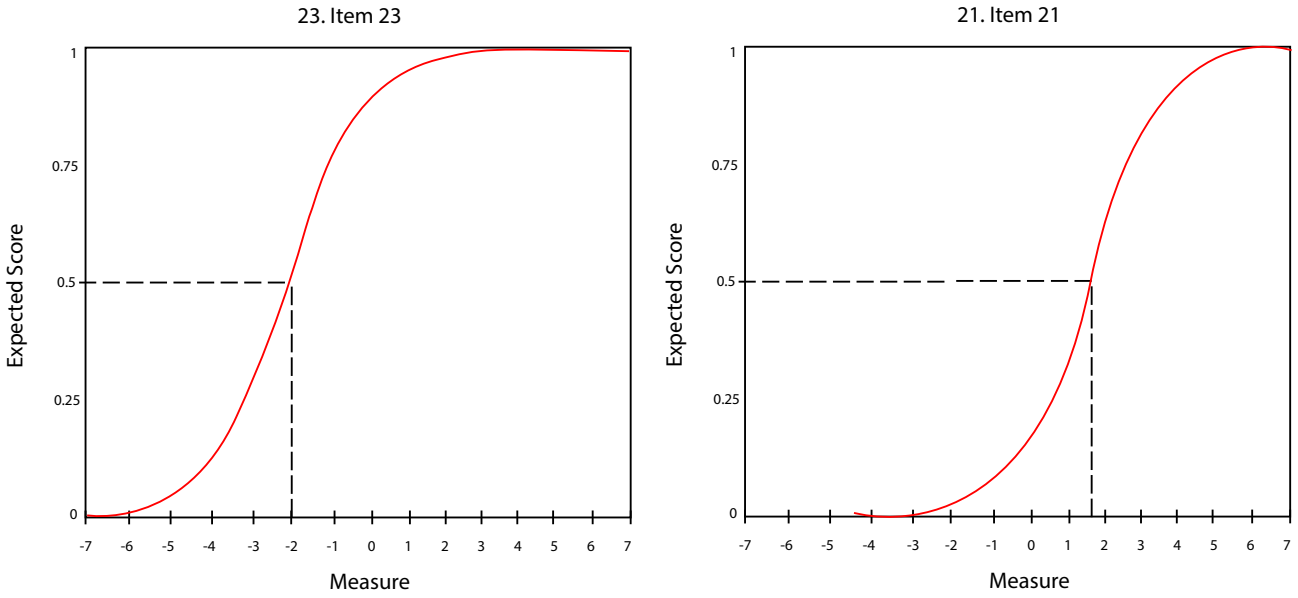

Figure 1:
Item Characteristic Curves (or ICCs) for item 21 and item 23

that 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. The positive fit statistics values indicate that the response string has more variation than expected, which implies that 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. The analysis of the items' standardized mean square values indicated a variation of 13 to 24 percent between the observed data and the response pattern predicted by the model. This discrepancy exceeds what is expected if the data and the model are completely compatible. The presence of these underfit items in the test has the potential to compromise its quality. Therefore, it is advisable for the test developer to thoroughly re-evaluate these items and scrutinize the construction process for any potential errors or mistakes.

Figure 2 illustrates the comparison between the observed student performances, depicted as the jagged empirical ICC (blue curve), and the Rasch model expectations, represented by the theoretical ICC (red curve), for item 13. This particular test item displays excessive variation and a near-perfect difficulty estimate of 0 logits.

Item	Raw Score	Measure	Model S.E.	INFIT		OUTFIT		MISFIT TYPE
				Mean Square	tz	Mean Square	tz	
8	84	.62	.14	1.16	2.4	1.18	2.2	Underfit
5	87	.56	.13	1.13	2.1	1.14	1.8	Underfit
9	88	.55	.13	.86	-2.4	.84	-2.3	Overfit
7	120	.02	.12	.87	-3.4	.85	-3.2	Overfit
13	121	.01	.12	1.24	5.5	1.32	6.0	Underfit
28	145	-.35	.12	1.15	4.2	1.18	4.0	Underfit
2	165	-.65	.12	.91	-2.9	.88	-2.7	Overfit
1	185	-.95	.12	.90	-2.7	.85	-2.7	Overfit
36	196	-1.12	.13	.90	-2.4	.91	-1.4	Overfit

Table 3:
Item Statistics: Misfit Measures

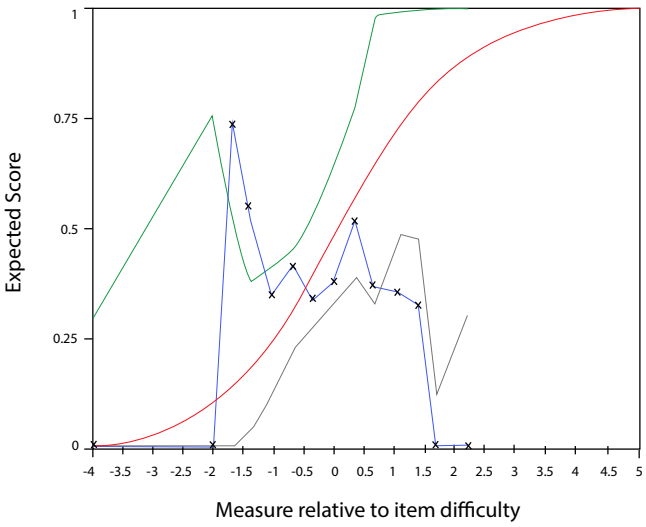

Figure 2.
Actual person performance versus theoretical item characteristics curve (ICC) for underfit item 13.

The output segment for item 13 reveals suboptimal fit characteristics when compared to the modeled expectations, despite its relatively acceptable mean square value. Poor fit statistics suggest that the actual test-takers' performances differ significantly from the modeled expectation. Specifically, a slight deviation was observed between -2.0 and -1.0 logits and from logits measures greater than 0.5. Additionally, the curve depicted in the analysis indicates that individuals with a lower level of ability are more likely to answer this item correctly than those with a higher level of ability, which runs contrary to the anticipated pattern. Items 7, 2, 1, 9, and 36, conversely, demonstrate overfitting of the model as their infit t values are lower than -2 logits and their outfit statistic is less than 0.75 logits. These values suggest a smaller degree of variation than what was modeled, indicating that the response pattern is more akin to a Guttman-style response string, where easy items are always correct, and difficult items are always incorrect. This conclusion is supported by the infit mean square value, which indicates a range of 9 to 14 percent less variation in the observed response pattern than expected. Although the presence of overfitting items has minimal practical implications, test developers should exercise caution because these items may overestimate the test item reliability, leading to an inflated estimation of the quality of the measure. Furthermore, omitting overfitting items may deprive the test of its best items. Thus, it is essential to revise these items before considering their removal from the test.

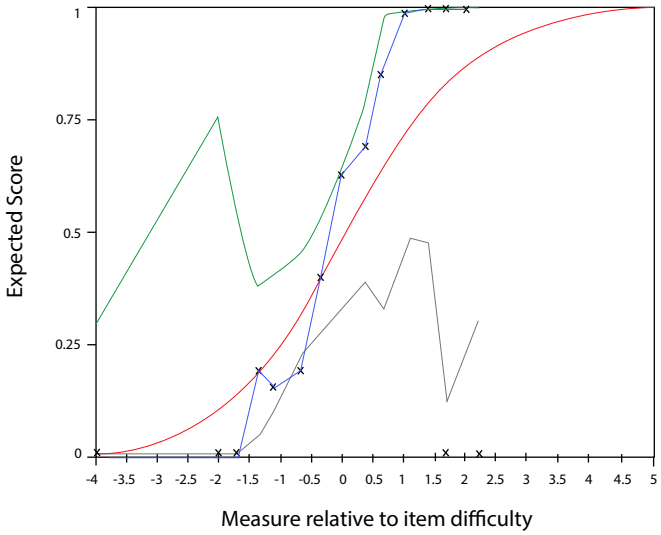

Figure 3.
Actual person performance versus theoretical ICC for overfit item 7

Figure 3 illustrates the comparison between the students' actual performance and the theoretical item characteristic curve (ICC) for Item 7, a test item that displays insufficient variation from the expected model. The graph shows that the observed performance of the students ranged from 1.0 to 2.0 logits higher than the predicted values. Similarly, their scores below -1.5 logits deviated from the model, indicating lower than expected performance. The overfitting of a low-ability group indicates that the item can distinguish between small differences in ability, which is considered desirable in Classical Test Theory. However, according to Rasch theory, it may indicate the presence of other factors, such as item dependency (Linacre, 2017). The item's level of difficulty was situated at the midpoint of the test, as displayed in Table 2. The mean square values for this item were not overly negative and were closer to the expected value of 1.0. Conversely, the t-statistics deviated farther from the model's anticipated values. In this case, the item adhered to the Guttman style, resulting in an overestimation of expectations based on the item's difficulty level and the students' abilities.

The findings on underfit and overfit items reveal that the data collected from these items do not conform to the Rasch Model, despite having a relatively accurate estimation of their level of difficulty. This is indicated by their outfit and infit t statistics falling outside the acceptable range. This suggests that these items are less compatible with the model than expected and may imply the presence of multidimensionality. Therefore, they should either be modified, discarded, or amended to focus on the target latent trait being tested. The assessment of well-fitting items against the Rasch Model serves to ensure the quality of the measurement instrument (Boone, Staver, and Yale, 2014). Furthermore, while the mean of unstandardized fit estimates (mean squares) is close to the expected value of 1, the mean and standard deviation of the standardized version of fit estimates (t statistics) show a slight deviation from the anticipated values of 0 and 1, respectively. These findings confirm that the test is less compatible with the model's expectations due to these misfit items. On the other hand, items that meet the criteria for the mean square value and t statistics are considered to have a good fit and are compatible with the expected model. Figure 4 displays the actual performance of students versus the theoretical item characteristic curve (ICC) for item 50, which is one of the well-fitted items.

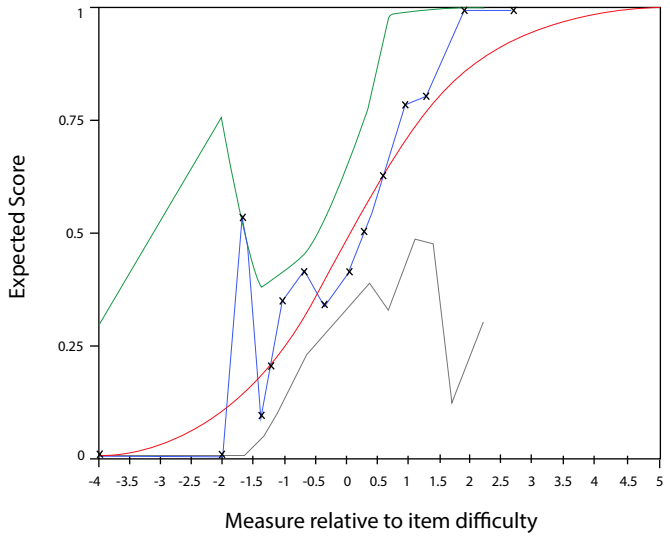

Figure 4:
Actual person performance versus theoretical ICC for item 50

The jagged curve (blue curve) represents the actual test performance of the 300 students, while the smooth curve (red curve) models the expected performance of the interaction between persons and the item. However, a perfect fit to the Rasch model is an unrealistic expectation. As demonstrated in Figure 4, the students' performance in item 50 is relatively close to the expected performance reflected by the Rasch model (the ICC), as evidenced by the plotted points of their mean actual responses. Although there was a slight deviation from the modeled curve around -2.0 logits concerning item difficulty, this variation is predicted by the Rasch model in terms of the actual deviation from the expected response. Additionally, the infit and outfit mean square values for both items were approximately 1.0, and their standardized versions, the infit and outfit t-statistics, were close to zero. These values indicate the compatibility of the item with the model, in addition to its difficulty measure estimate, which is located at the midpoint of the test with an almost exact value (0 logits).

Item Difficulty and Student Ability Relations

Figure 5 presents a Wright map that illustrates the relationship between the difficulty of the items and the ability of the students. As a quality indicator, this graphical representation connects the item difficulties and student ability estimates on a common scale to ensure that both variables are aligned to maximize the test's informativeness (Junpeng, 2020). The distance of the step from the bottom of the path on the map represents the item's difficulty relative to other items, providing a representation of the item difficulty. Closer to the bottom indicates easier items, while farther away denotes more challenging items. Based on the representation of persons and items on the map, Item 33 was much more difficult than Item 23, and Item 21 was the most challenging item in this test. Most students did not perform well on Item 21, whereas Item 23 was the easiest, with the majority of students answering it correctly. These findings demonstrate that all items were useful for discriminating ability among the students in this group, as not everyone was successful on the easiest item or got the most difficult item wrong. Furthermore, items 7, 13, 45, and 50 are located

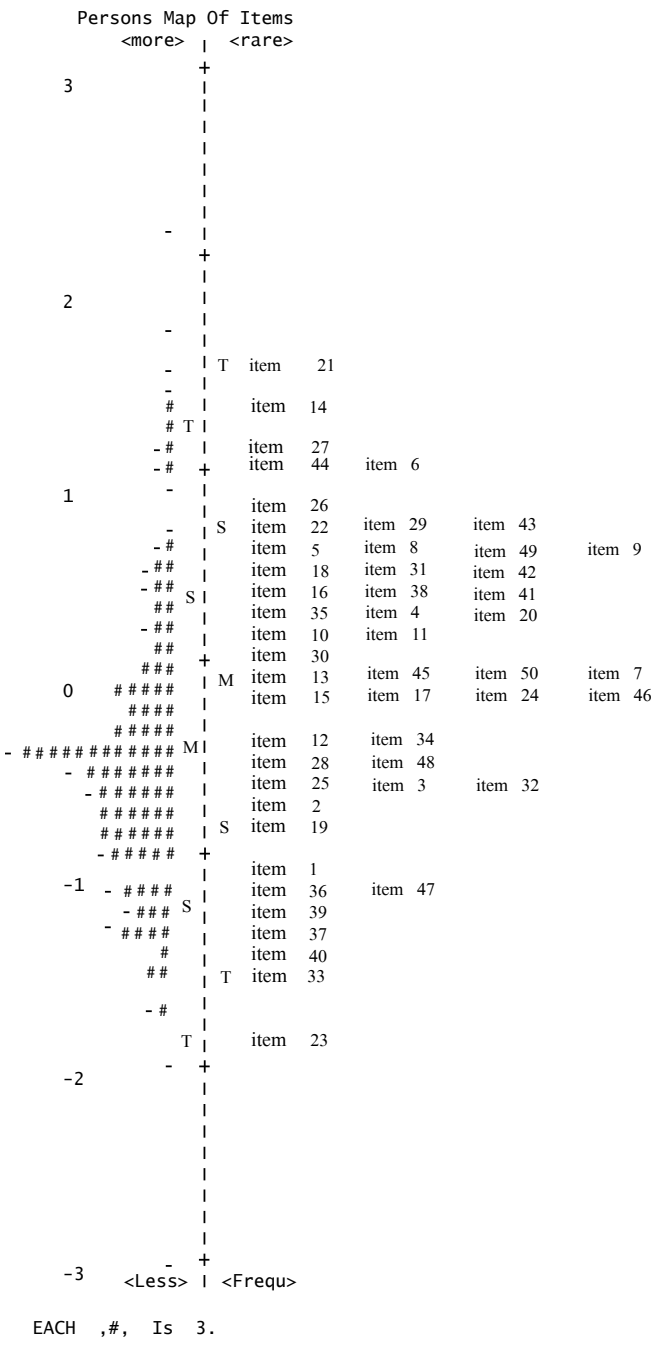

Figure 5:
Item-person Wright Map

at 0 points in the item-person map, indicating an almost exact difficulty estimate of 0 logits. Nine students had a 50 percent chance of getting these items correct, and an additional 57 students could potentially answer these items correctly with more than a 50 percent probability of success. However, the remaining 78 percent of test takers failed to answer these items correctly. These results suggest that this type of test was somewhat difficult for the examinees' level of ability, despite two students achieving a perfect score of 50. Only one-third of the total samples had an equal or higher than 50 percent chance of obtaining a correct answer on half of the total items. This indicates that the majority of examinees' ability levels did not surpass the difficulty level of the majority of the items. The figure shows that the majority of examinees failed the exam, which could be considered a significant inadequacy in a general test development perspective. To address this issue, the test may require even easier questions to raise the "ceiling" of the test or teachers may need to implement a better teaching strategy to facilitate their students' learning (Bond & Fox, 2012).

Item and Person Reliability

Table 4 presents the reliability of the test items, providing an overall summary of the mean difficulty estimates of the items, item and person reliability, and separation. The table reveals a large positive value for the item difficulty mean estimate, indicating that the test was challenging for the sample group of students who took the examination, which is consistent with the results depicted in the item-person Wright map shown in Figure 5. The standard deviation of 46.1 for item estimates suggests a greater spread of item measures or variation in those measures than with person measures. Regarding item reliability, this test exhibited a very high degree of reliability with a score of 0.97 on a scale of 0 to 1. Moreover, the item separation value of 5.54 indicates that the persons have differentiated more than five levels of item difficulty. These findings suggest that we can rely on this order of item estimates to be replicated when administering the same test to other suitable groups of students. 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. Further examination of the effects of these items on test reliability should be conducted. However, the person reliability index of 0.77 is relatively high, indicating that if the same

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 4.
Summary of Item and Person Estimates

group of persons were given another set of items measuring the same construct, almost the same estimate of a person's ability would be expected. The person separation value of 1.81 suggests that the items were able to differentiate between more than one level of a person's ability.

Unidimensionality of the Test

To examine the unidimensionality of the test, Principal Component Analysis (PCA) of residuals in Rasch was conducted, as a unidimensional test measures only one's ability (Susongko, 2016). Tables 5 and 6 summarize the findings of this analysis. Table 5 shows that the observations had a total variance of 65.9 eigenvalue units. Out of this total variance, 15.9 eigenvalue units were explained by person and item measures. However, the unexplained variance had 50 eigenvalue units, covering over 75 percent of the total variance. This significant difference from the Rasch measure indicates that this unexplained variance may have arisen from sources not intended to be included in the test, and therefore, was not accounted for by the Rasch measurement. The results suggest that the unexplained variance may be attributed to substantive structures. Following the Rasch measurement, the residuals were observed, as reflected in the contrasts in the findings. Residuals refer to the difference between students' observed performance on an item and what is expected by the Rasch model. A smaller residual indicates a better fit of the data to the model (Kazemi, 2020). In this case, the results indicate the presence of five contrasts, some of which consist of more than two eigenvalue units, suggesting the possibility of a potential dimension (Linacre, 1998; Ee, Yeo, and Kosnin, 2018). Consequently, a Principal Component Analysis of Rasch residuals (PCAR), or linearized Rasch residuals, was conducted to extract meaningful information from these contrasts. Table 5 summarizes the findings of the first factor, which had the highest factor sensitivity ratio among the contrasts.

The results of the analysis revealed that the first contrast had an unexplained variance of 3.2 units, indicating that approximately three eigenvalues contributed to a subdimension in the data. These items shared a common characteristic beyond the Rasch dimension, leading to their clustering together. Figure 6 illustrates the factor plot of the standardized residuals after extracting the primary Rasch dimension. The plot displays higher factor loadings for items 21(A), 44(B), and 9(C), which are located at the top of the map. These items have significant variance that remains unexplained by the primary Rasch measure. Table 6 presents the factor loadings for the first dimension (contrast 1), which reveal three items (21, 44, and 9) with substantial positive loadings on the factor found in the item residuals (i.e., with an off-dimension loading of 0.4 or greater). In contrast, two items (40 and 36) exhibited negative correlations with the factor.

		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 5.
Table of Standardized Residual Variance (in Eigenvalue Units)

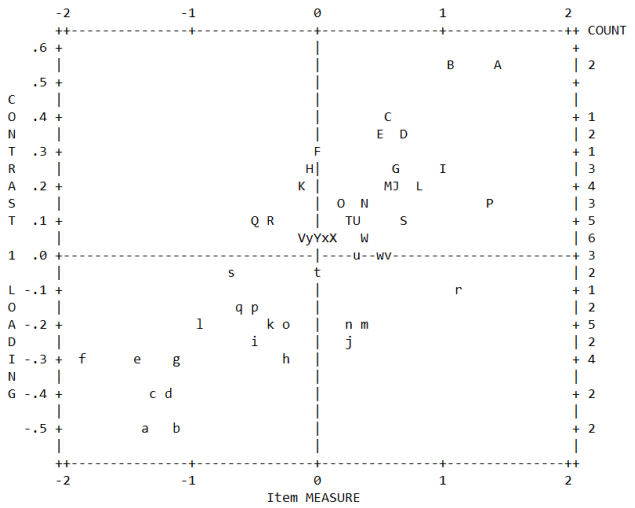

Figure 6.
Standardized Residual Variance Scree Plot (Contrast 1)

These findings provide empirical evidence of a separate subscale. However, it is up to the researcher to determine whether this subscale is significant and large enough to be measured separately from the Rasch measures. The researcher must weigh the costs and benefits of including these items as part of the original Rasch dimension, which may result in a loss of sensitivity or validity of the measurement, or excluding these items from the total score and working towards assessing and interpreting the other dimension separately. teaching strategies to facilitate the students' learning. 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.

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
-.50	-1.37	.95	.98	a 40	Item 40
-.50	-1.12	.90	.91	b 36	Item 36

Table 6.
Principal Component Analysis of Standardized Residual Correlations for Items on First Dimension (Sorted By Loading)

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

References

Baghaei, P. (2008). The rasch model as a construct validity tool. Transactions of the Rasch Measurement SIG Amer-

ican Educational Research Association, 22(1). ISSN 1051- 0796

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 https://doi.org/10.1187/cbe.16-04-0148

Boone, W., Staver, JR., & Yale, MS. (2014). Rasch Analysis in the Human Sciences. Springer https://doi.org/10.1007/978-94-007-6857-4

Chan, S., Ismail, Z., & Sumintono B. (2014). A rasch model analysis on secondary students' statistical reasoning ability in descriptive statistics. Procedia - Social and Behavioral Sciences, 129,133-139. https://doi.org/10.1016/j.sbspro.2014.03.658 https://doi.org/10.1016/j.sbspro.2014.03.658

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

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 https://doi.org/10.1016/S0187-893X(13)72494-7

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), https://doi.org/10.11591/ijere.v7i4.15376

264-269. http://doi.org/10.11591/ijere.v7i4.15376 https://doi.org/10.11591/ijere.v7i4.15376

Fatimah, S., Elzamzami, A. B., & Slamet, J. (2020). Item Analysis of Final Test for the 9th Grade Students of SMPN 44 Surabaya in the Academic Year of 2019/2020. JournEEL (Journal of English Education and Literature), 2(1), 34-46. https://doi.org/10.51836/journeel.v2i1.81 https://doi.org/10.51836/journeel.v2i1.81

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 https://doi.org/10.11591/ijere.v9i3.20503

Kantahan, S., et al. (2020). Designing and verifying a tool for diagnosing scientific misconceptions in genetics topic. International Journal of Evaluation and Research in Education (IJERE), 9(3), 564-571. http://doi.org/10.11591/ijere.v9i3.20544 https://doi.org/10.11591/ijere.v9i3.20544

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 https://doi.org/10.46504/14201805ka

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

Linacre, J. (1994). Sample size and item calibration stability. Rasch Measurement Transactions, 7(4), 328

Linacre, J. (2012). Winsteps help for Rasch analysis

Mokshein, S., Ishak, H., & Ahmad, H. (2019). The use of rasch measurement model in english testing. Cakrawala Pendidikan, 38(1). https://dx.doi.org/10.21831/cp.v38i1.22750 https://doi.org/10.21831/cp.v38i1.22750

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 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

Souza, M.P., et al. (2017). Rasch analysis of the participation scale (P-scale): usefulness of the p- scale to a rehabilitation services network. BMC Public Health, 17(1), 934. https://doi.org/10.1186/s12889-017-4945-9 https://doi.org/10.1186/s12889-017-4945-9

Sumintono, B. (2018). Rasch model measurements as tools in assessment for learning. Advances in Social Science, Education and Humanities Research, 173, 38-42. https://doi.org/10.2991/icei-17.2018.11 https://doi.org/10.2991/icei-17.2018.11

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

Taber, K. (2018). The use of cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education, 48(1), 1273- 1296. https://doi.org/10.1007/s11165-016-9602-2 https://doi.org/10.1007/s11165-016-9602-2

Schumacker, R. (2004). Rasch measurement: the dichotomous model. Introduction to Rasch Measurement: Theory, Models and Applications, 226-253

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 https://doi.org/10.1155/2018/8719012

Winarti, A., Almubarak, Annur, S. (2019). How does the rasch model justify multiple choice question items as a measure of student understanding of acid-base material at the sub- microscopic level. International Journal of Innovation, Creativity and Change, 7(11). https://repo-dosen.ulm.ac.id//handle/123456789/13967

Wright, B. & Linacre, J. (2002). Understanding rasch measurement: Construction of measures from many faced data. Journal of Applied Measurement, 3(4), 486-512