Jurnal Penelitian dan Evaluasi Pendidikan
Volume 28, No. 1, June 2024 (79-93)
Online: http://journal.uny.ac.id/index.php/jpep

Individual ability on high-stakes test: Choosing cumulative score or rasch for scoring model

Muhammad Dhiyaul Khair*, Sukaesi Marianti
Universitas Brawijaya, Indonesia
*Corresponding Author. E-mail: khairdhiyaul@gmail.com

ARTICLE INFO

Article History
Submitted:
01 March 2024
Revised:
04 June 2024
Accepted:
08 June 2024

Keywords
Ability; cumulative score; Rasch, dichotomous responses

Scan Me:

ABSTRACT

In a test, a method is required to estimate an individual's ability based on their responses. Typically, this is done by summing the correct responses or calculating a cumulative score. An alternative method is the Rasch model. This study aims to determine whether an individual's position, based on cumulative score estimates, remains unchanged or changes when compared with ability estimates using Rasch on dichotomous responses. The study uses open-source data from the 2018 Program for International Student Assessment (PISA) by the Organization for Economic Co-operation and Development (OECD) and involves 317 Indonesian students. Ability analysis will be conducted on Math and Reading aspects using cumulative scores and Rasch with dichotomous responses. The study will employ data analysis techniques such as Rasch, paired samples t-test, and descriptive statistical analysis. The cumulative score and Rasch results will be tested using a paired samples t-test, and a comparison of the cumulative score and Rasch estimation results will be carried out using descriptive statistical analysis. The study results indicate that there are differences in individual positions based on ability estimates using cumulative score and Rasch. These differences are caused by variations in scores. Therefore, even if two individuals have the same cumulative score, they may have different Rasch estimates.

To cite this article (in APA style):
Khair, M. D., & Marianti, S.(2024). Individual ability on high-stakes test: Choosing cumulative score or rasch for scoring model. *Jurnal Penelitian dan Evaluasi Pendidikan, 28(1)*, 79-93 doi: https://doi.org/10.21831/pep.v28i1.71661

INTRODUCTION

Currently, there are numerous methods and tools available for measuring an individual's abilities. One popular method in society is through the use of tests. A test is a series of items designed to measure the extent of an individual's ability or to discover a specific aspect of their abilities (Widoyoko, 2012). A reliable test is one that can provide accurate information and data to represent an individual's true abilities (Saifuddin, 2002; Sarea & Ruslan, 2019). Tests are typically categorized as either cognitive or non-cognitive.

Cognitive tests measure an individual's potential or achievement. An example of a cognitive test is the Program for International Student Assessment (PISA). PISA measures individual abilities or capabilities as a reference for learning evaluations carried out in a country (Kemendikbud, 2019). Cognitive test items come in various forms, such as multiple-choice, true/false, and open-ended questions (OECD, 2017). These items categorize individual responses as either correct or incorrect. Tests with dichotomous responses require an approach to estimate the ability being measured. Researchers typically use Classical Test Theory or Item Response Theory. Both approaches allow for the estimation of the ability being measured.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v28i1.71661

Classical Test Theory (CTT) is a widely used model for accurately estimating abilities (Fernanda & Hidayah, 2020). CTT introduces three concepts: observed score, true score, and error (Bichi, 2016). The true score is the score that a test respondent would receive if there were no errors in the measuring instrument. However, this is highly unlikely as tests are rarely perfect. Therefore, the observed score for each respondent is influenced by error, either higher or lower (Vincent & Shanmugam, 2020). The focus of CTT is on the total test score (Magno, 2009). The cumulative score, also known as the total score, is calculated by adding up the number of correct item responses for each individual. This score is used to estimate the ability of CTT. In the case of dichotomous responses, a correct response is given a score of 1 and an incorrect response is given a score of 0 (Amelia & Kriswantoro, 2017).

Although applying CTT is relatively straightforward, it presents several challenges. According to Rusch et al. (2017), CTT has three limitations. For instance, CTT assumes a linear relationship between latent variables and cumulative scores, which rarely represent behavioural constructs in real-life situations. Additionally, the true score cannot be estimated directly; rather, it can be estimated using assumptions that are difficult to fulfil, and the parameters depend on the sample. Furthermore, the cumulative score assumes equal weighting for all items without any consideration for their individual importance. Additionally, a significant drawback of CTT is its dependence on tests, which means that the difficulty level of the test can directly impact the resulting scores (Bichi, 2016).

To address these limitations, previous researchers have proposed a model known as Item Response Theory (IRT) (Maulani & Rahardjo, 2014). The item response theory (IRT) approach is used to measure the likelihood of a test respondent answering an item correctly or incorrectly based on item analysis. Each test item has an item characteristic curve that describes the probability of a correct or incorrect response based on ability estimates. The IRT approach considers three important parameters: difficulty, discrimination, and guessing (Fan, 1998).

One model that uses IRT is Rasch. Rasch is a specialized IRT model that focuses on basic measurement requirements and is relatively easy to understand. In contrast, IRT, in general, is concerned with fitting flexible models to observed data (Rusch et al., 2017). Rasch focuses on one of the parameters used in IRT, difficulty (Fan, 1998). Rasch is a unidimensional probabilistic model that states that the easier a question is, the more likely it is that the respondent will answer it correctly, and the higher the respondent's ability, the more likely they are to answer the question correctly compared to respondents with low ability (Magno, 2009).

IRT has several advantages: it assumes non-linear relationships, allows for more accurate true score estimation, can estimate item parameters independent of the sample, and allows the researcher to select items that fit the desired model (Rusch et al., 2017). In addition, Rasch provides a methodology that allows the examination of hierarchical structure, unidimensionality and measurement additivity (Prieto et al., 2003). Based on a study by Magno (2009), Rasch's estimates of difficulty do not change across samples compared to CTT, which is inconsistent, and difficulty is more stable across test forms compared to the CTT approach.

On the other hand, Rasch is a very strict model because it requires only one latent variable underlying the test, and all items have the same discrimination parameters (Rusch et al., 2017). Because it is a very strict model, Rasch is not flexible enough to be useful for modelling. Violations of IRT model assumptions or discrepancies between the IRT model used, and the test data can lead to incorrect or unstable IRT model parameter estimates. This is because, when applying any IRT model, it is important to assess the extent to which the assumptions of the IRT model are valid for the particular data and how well the test data fit the chosen model in the particular situation (Fan, 1998).

Studies comparing CTT and IRT have found that CTT and IRT produce similar items and abilities when compared (Fan, 1998). This statement is also supported by a more recent study that shows similarities between CTT and IRT in terms of the estimation of items and abilities (MacDonald & Paunonen, 2002). However, a later study found that CTT ability estimates are invariant across different item sets, whereas IRT ability is more invariant across conditions (Progar & Sočan, 2008). In contrast, Xu & Stone's (2012) study showed that the cumulative score was slightly better than the IRT-based score for a short scale (10 items) and a small sample (N=250).

Based on the explanation from the previous study above, CTT and IRT serve as approaches to accurately estimate individual abilities. Although it is known that CTT and IRT each have advantages and disadvantages, previous studies have not explained in more detail the extent of the differences between cumulative score and Rasch in estimating individual ability. This raises the question of whether, when the test is measured using different measurement models, specifically cumulative score and Rasch, the estimated ability for each individual is the same or whether there are differences, as this will affect the accuracy of the test in estimating ability. Therefore, test measurements are carried out using at least two or more different measurement models so that the measurement models can be compared to see which measurement model is more appropriate for the test.

Thus, this study aims to compare the cumulative score measurement model of the CTT and the Rasch measurement model of the IRT. The cumulative score is obtained from the total individual score produced by the test, while Rasch is obtained from Rasch estimates based on the test scores. Once these two things were obtained, the cumulative score and Rasch were compared. Comparisons are made by looking at the t-test between the two and the extent to which each model provides information about the estimated ability of each individual.

RESEARCH METHOD

Cumulative Score

Classical Test Theory is considered a "true score theory", which assumes that differences between test takers' responses are systematic; they are influenced by variations in test takers' abilities (Vincent & Shanmugam, 2020). The main concept of CTT is that the observed score (X) consists of a true score (T) and an error score (E), where the true score and the error score are independent of each other (Magno, 2009). The true score is the actual score of an individual's ability or skill. However, the true score cannot be estimated directly; it is estimated from observed scores, which may change each time the test is administered. The thing that affects changes in the observed score is error.

Different models have been formulated based on this concept (Bichi, 2016). One of them is the "classical test model", which is formulated as follows:

$$X = T + E \tag{1}$$

10.21831/pep.v28i1.71661

This formula is a simple model that relates the observed score (X) to the sum of two unobservable variables: the true score (T) and the error score (E). Because true scores cannot be observed directly, true scores must be estimated from individual responses to a series of items in the test. Therefore, a number of simplifying assumptions must be made in order to solve the equation (Bichi, 2016).

In general, cumulative scores are used in scales or achievement tests to determine individual ability. The cumulative score is obtained from the sum of all correct items based on individual responses to each item (Kiliç, 2019). Correct responses are given a score of 1, while incorrect responses and no responses are given a score of 0. In cumulative scoring, it is assumed that there are no errors, so the cumulative score obtained by an individual represents the individual's abilities.

The limitation of the cumulative score is its dependence on the level of difficulty of each item in the test (MacDonald & Paunonen, 2002). According to Sarea and Ruslan (2019), a good test is one that has a proportional distribution of items; that is, the test has items of easy, medium, and hard so that the test is able to estimate abilities in both the low and high categories. For example, if the majority of the items in a test are of low difficulty, the results of the test will show that the test takers have produced a high category of cumulative scores. On the other hand, if the test has many items with a high level of difficulty, then the test takers will show a low category of cumulative scores.

Rasch Model

Item Response Theory is called strong true score theory or modern mental test theory because IRT is a newer set of theories and makes stronger assumptions than CTT (Magno, 2009). According to Hambleton & Jones (1993), IRT is a statistical theory about test performance, items performance, and how test performance is related to the abilities measured by the item in the test. Item responses may be discrete (dichotomous or polytomous) or continuous; item score categories may or may not be sorted; and there may be one ability or many abilities underlying test performance.

The basic concept of IRT is that item performance is related to the estimated number of latent traits of the respondent (Anastasi & Urbina, 2002; Magno, 2009). A latent trait (ability) is symbolized by theta (θ), which refers to a statistical construct (Magno, 2009). According to Abedalaziz & Leng (2018), there are two main concepts of IRT, namely that the performance of a test taker on a test item is a function of their traits or abilities; and the graphical relationship between the ability traits of test takers and their probability of answering an item correctly, in the form of a monotonically increasing function called the item characteristic curve (ICC). Since item performance depends on ability, as the level of ability increases, the probability of a correct response increases or remains the same (Abedalaziz & Leng, 2018; Hambleton et al., 1991).

Within IRT, there are different models that can be used to measure test performance. The application of each IRT model depends on the particular situation, namely based on the nature of the test items and the feasibility of the theoretical assumptions about the test items (Fan, 1998). Specifically for test items that have a dichotomous response (0 or 1), there are three IRT models known as three-parameter (3PL), two-parameter (2PL), and one-parameter (1PL) (Bichi & Talib, 2018). The parameters used in the model are difficulty, discrimination and guessing. In addition to these three models, there is a no less popular model, the Rasch model, which focuses on the difficulty parameter. For practical purposes, when each individual in the person sample is parameterized for item estimation, the resulting model is Rasch. Conversely, when the person sample is parameterized by a mean and standard deviation for item estimation, the resulting model is 1PL (Rasch, n.d.). However, Rasch and 1PL models are mathematically equivalent because values from one model can be transformed to the other by appropriate rescaling (Hayat et al., 2020).

The Rasch model is suitable for modelling dichotomous responses and models the probability of a respondent's correct response to a dichotomous item (Magno, 2009). This is based on the underlying logic that subjects have a higher probability of correctly answering easier items and a lower probability of correctly answering more difficult items (Columbia University, 2016). This model is called a prescriptive model because it sets specific conditions that must be met by the data. This means that from the beginning, the entire research process must be in line with the specifications of the model (Bichi et al., 2019). The Rasch model is based on the assumption that the estimation and discrimination parameters are negligible or constant (Magno, 2009). In this model, item discrimination is set to a value of a = 1 for all items, and only the item difficulty parameter can have a different value (Baker & Kim, 2017). Below are the equations used by the Rasch model:

$$P_i(\theta) = \left(\frac{1}{1+e^{-a(\theta-b)}}\right) = \left(\frac{1}{1+e^{-L}}\right), a = 1 \qquad (2)$$

In this equation, a is the discrimination parameter, because the discrimination parameter is ignored or constant, then a here becomes 1. Then, b is the difficulty parameter, θ (theta) represents ability, e is a constant of 2.718, and L is the logit deviation which consists of $a(\theta - b)$. From the explanation of this equation, this will get the probability that the respondent answered the item correctly for each item in the test.

Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is used to obtain ability estimates in IRT, including Rasch. The estimation starts with the a priori value of the respondent's ability and the item parameter values. These values are used to estimate the probability of a correct response to each item. This process is repeated until the change in the estimated ability is negligible. The result of this process is an estimate of the respondent's ability (Baker & Kim, 2017). Ability estimates based on the Rasch model can be formulated as follows:

$$\theta_{s+1} = \theta_s - \frac{\sum_{i=1}^{I} 1[u_i - P_i(\theta_s)]}{-\sum_{i=1}^{I} 1 P_i(\theta_s) Q_i(\theta_s)} \qquad (3)$$

$$Q_i(\theta_s) = 1 - P_i(\theta_s) \qquad (4)$$

θs is the estimated ability of the test taker in *s* repetitions; u_i is the response of the test taker to item *i*, where $u_i = 1$ for a correct response and $u_i = 0$ for an incorrect response. $P_i(\theta_s)$ is the probability of a correct response for item *i*, based on the results of item characteristic curve model, at ability θ in *s* repetitions. $Q_i(\theta_s) = 1 - P_i(\theta_s)$ is the probability of an incorrect response for item *i*, based on the results of item characteristic curve model, at ability θ in *s* repetitions.

MLE is unable to estimate ability in several conditions. First, when the test taker does not answer all items correctly, the estimate will result in an infinitely negative score. Second, when the test taker answers all items correctly, the estimate will produce an infinitely positive score. In both cases, it is impossible to obtain an ability estimate for the subject. As a result, programs must have certain procedures to avoid both of these conditions. When the program finds a test score of zero or a perfect test score, the program will eliminate the subject from further analysis (Baker & Kim, 2017).

Data

This study uses dichotomous response data from previous research. The data were obtained from the Programme for International Student Assessment 2018, or PISA 2018, organized by the Organisation for Economic Co-operation and Development (OECD, 2019). PISA is a test that measures reading, mathematics and science skills. The purpose of PISA is to assess the education system in a country that is a member of the OECD. PISA 2018 was conducted in 79 countries, one of which was Indonesia, and it was conducted from 19 March 2018 to 19 April 2018 (Kemendikbud, 2019).

The data subjects used in this study were 317 out of 12,098 students in grades 7 to 12 from 397 schools in Indonesia (OECD, 2019). The aspects of PISA used are Math and Reading. The number of items used in this study in the Math aspect was 20 items, and in the Reading aspect was 22 items. The responses from the two aspects of the data were in the form of dichotomous answers, namely wrong answers (0) and right answers (1). The decision on the number of samples for this study was based on Linacre's (1994) statement that Rasch's research with a sample of 300 subjects and 20 items resulted in a 99% confidence interval.

Method

This study uses quantitative psychometric research methods. This study compares different test measurement models, using classical and modern test measurement models, using the CTT and the Rasch paradigm. The two models are each analyzed, and then the results of the two analyses are compared.

The first step is to find data that meets the needs. Once the data are obtained, adjustments and data cleaning are carried out, such as checking for empty data and checking for inappropriate items. The analysis used is cumulative score and Rasch analysis. From the results of these two analyses, a t-test was carried out to see if there was a difference between the two. Finally, descriptive statistical analysis was used, which aims to describe the results of the analysis in terms of statistical techniques. The focus of this study is to determine whether the position of individuals based on ability estimation using cumulative score remains the same or changes when compared to ability estimation using Rasch on dichotomous responses.

Rasch

The first data analysis technique was Rasch. This technique was applied using the "eRm" package in the R program. Rasch computation produces difficulty and theta (θ) parameters.

Cumulative Score

The cumulative score was also analyzed using the R program. The resulting parameter is an estimate of individual ability based on the sum of correct answers on the test.

Paired Samples T-test

One of the techniques used in this study is the t-test technique, specifically the paired samples t-test. The t-test analysis is used to compare the means between the cumulative score and the Rasch theta in both Math and Reading. The first group was the cumulative score and the second group was the Rasch theta. The purpose is to see if there is a difference between the cumulative score and the Rasch analysis results. Before the t-test is performed, the cumulative score is converted to z-score so that the mean and standard deviation are standardized to 0 for the mean and 1 for the SD. This is to ensure that the scales of the cumulative score and the Rasch theta are equivalent.

Statistical Analysis Techniques

The techniques used in descriptive statistical analysis are histograms and density plots. Histograms are used to show the frequency and distribution of ability at a given level from both cumulative score and Rasch results. When the ability distribution is visible, it can be compared with the cumulative score histogram and the Rasch histogram. The density plot is used to visualize the difference between the cumulative score and Rasch

FINDINGS AND DISCUSSION

Findings

Cumulative score is one way of estimating the subject's ability from CTT by summing the correct answers for each subject. Cumulative scores obtained from Math and Reading datasets are visualized using the histograms shown in Figure 1a and Figure 2a. In Figure 1a, the Math dataset shows that the range of cumulative scores obtained by the subjects is from 0 to 19, with the highest frequency at score 3, i.e. 53 people,, while the lowest frequency is at scores 16 and 19, i.e. 1 person each. Then, Figure 2a shows that the range of cumulative score in the Reading dataset is between 8 to 22, with the two highest frequencies being at score 20 with a total of 40 people and score 21 with a total of 39 people, while the lowest frequencies are at scores 8 and 9 with a total of 3 people each.

The method used by Rasch to estimate subject ability is Maximum Likelihood Estimation or MLE. First, the data set is computed using Rasch to produce difficulty parameters for each item. Then, the difficulty parameter is calculated again using MLE, resulting in an estimate of each subject's ability. This Rasch ability estimate is called Theta (θ). Just like the cumulative score, Theta in Math and Reading are visualized with histograms, namely in Figures 1b and 2b. In Figure 1b, the range of Math theta shown is from -4 to 4, with the highest frequency at theta -2.0989, which corresponds to 36 people. In contrast to Math, Figure 2b shows that Reading's theta range is between -1 and 4. Theta 3.3939 is the Theta with the highest frequency in the Reading dataset, 38 people, followed by theta 2.6114 and 4.246 with 37 people each, while the lowest frequency is achieved by one person at 20 different theta points.

Table 1. Score variations based on computation of cumulative score and Rasch

Aspect	Variations	
	Cumulative Score	Theta
Math	19	72
Reading	15	50
Total	34	122

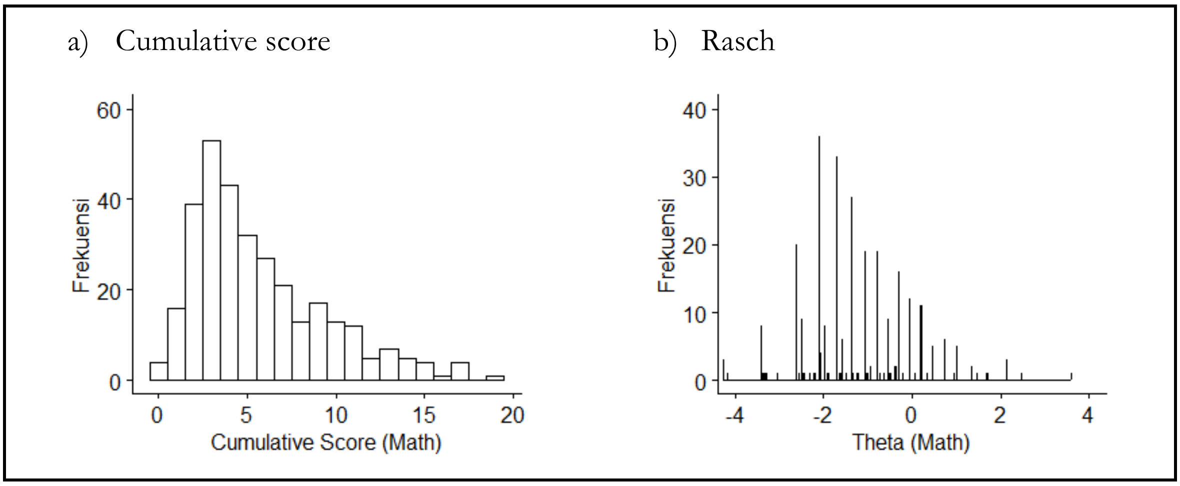
Comparing Between Cumulative Score and Rasch

Math

Figure 1 shows a visualization of the comparison between cumulative score and theta in Math in the form of a histogram. In Figure 1, the visible difference is the difference in the range between the cumulative score and theta, as the cumulative score range is between 0 and 19, while the theta range is between -4 and 4. The cumulative score was not standardized for this comparison before the t-test. This was done for two reasons: firstly, to facilitate visual comparison of the data, and secondly, to provide evidence as to why there are differences between the cumulative score and theta. Another difference is the total score variation. Based

10.21831/pep.v28i1.71661

on Table 1, this is because the variation in the cumulative score is less than the variation in theta score, which is 19 variations for cumulative score and 72 variations for theta.



a) Cumulative score

b) Rasch

Figure 1. Ability estimation histograms on Math

A more in-depth analysis of the difference in ability estimation between cumulative score and theta is shown in Table 2, which presents the variation of theta score and its response pattern based on subjects who get a cumulative score of 3. Based on Table 2, there are 53 subjects who get a cumulative score of 3. When estimated with theta, these 53 people are divided into six groups because when estimated with theta, the scores show differences. The reason for the difference in theta is that each group has a different NA response pattern. For example, theta -2.0989 has no NA response, and theta -2.0759 has an NA response in item M8. In addition, the NA response position also affects the theta score, which is exemplified in theta -2.0759, with the NA response in item M8 being lower than theta -1.9794 with the NA response in item M3. This is due to the difference in difficulty parameters in items M8 and M3. Looking at the difficulty parameter, the difficulty or beta of item M8 (.8102) is greater than the beta of item M3 (-1.0453).

Table 2. Variation of theta score based on cumulative score 3 in Math

Theta	Response Pattern	Frequency
-2.0989	x x x x x x x x x x x x x x x x x x x x	36
-2.0759	x x x x x x x NA x x x x x x x x x x x x	4
-1.9794	x x NA x x x x x x x x x x x x x x x x x	8
-1.9248	x x x x x x x x x x x x NA x x x x x x x	1
-1.8817	NA x x x x x x x x x x x x x x x x x x x	1
-1.7116	NA x NA x x x x NA x x x x x x x x x x x x	2
	Total	53

Note. x = correct or incorrect response, NA = missing response

Another finding is that as long as the cumulative score and NA items are the same, the theta obtained by the subject remains the same, even though the subject answers the correct response at different items. This statement is evidenced in Table 3 which presents examples of subjects and theta scores obtained based on cumulative score 3. In Table 3, it can be seen that subject 16, who answered the correct responses in items M10, M13, and M20 and subject 28, who answered the correct responses in items M8, M10, and M19, both got theta -2.0989.

10.21831/pep.v28i1.71661

Likewise, subject 190, with correct item patterns in items M9, M19, and M20 and subject 249, with correct item patterns in items M6, M13, and M14, get the same theta score of -1.7116.

Table 3. Example of subject and theta based on cumulative score 3 in Math

Subject	Theta	Response Pattern
16	-2.0989	0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1
28	-2.0989	0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0
190	-1.7116	NA 0 NA 0 0 0 0 NA 1 0 0 0 0 0 0 0 0 0 1 1
249	-1.7116	NA 0 NA 0 0 1 0 NA 0 0 0 0 1 1 0 0 0 0 0 0

Note. 0 = incorrect response, 1 = correct response, NA = missing response

Reading

Figure 2 shows the histogram of the cumulative score and theta in reading. The cumulative score ranges from 8 to 22, while the theta score ranges from -1 to 4. The total variation of cumulative score is 15, and the total variation of theta score is 50. These results indicate that there are differences in the range and total score variation between the two methods in both Reading and Math.

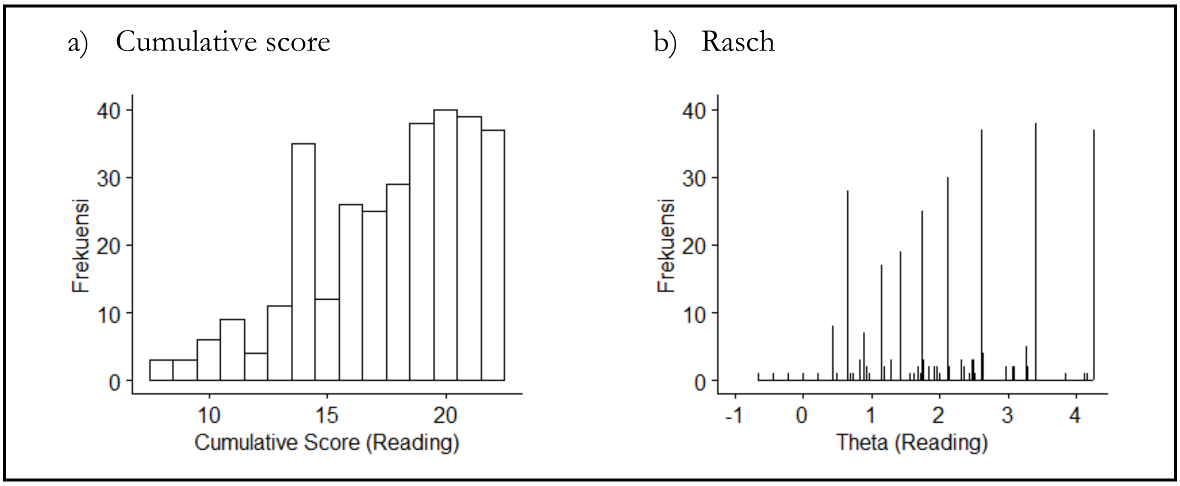


Figure 2. Ability estimation histograms on Reading

Table 4 shows the variation of theta scores for subjects who achieved a cumulative score of 20 in Reading. A total of 40 subjects achieved this score. The theta score for these subjects resulted in three different variations of theta. This variation is caused by the same factor as in Math, namely the difference in NA response patterns. The difference between theta 2.6114 and theta 3.2884 is that the former has a response pattern without NA, while the latter has an NA response pattern in item R20. Another finding is that subjects with the same cumulative score but with a greater number of NA responses show a higher theta than those with fewer NA responses. This is demonstrated by theta 2.6114, where there were no NA responses, resulting in a lower theta than theta 4.1213, where there were two NA responses.

10.21831/pep.v28i1.71661

Table 4. Variation of theta score based on cumulative score 20 in Reading

Theta	Response Pattern	Frequency
2.6114	x x x x x x x x x x x x x x x x x x x x x x	37
3.2884	x x x x x x x x x x x x x x x x x x x x x NA	2
4.1213	x x x x x x x x x x x x x x x x x x x x NA NA	1
	Total	40

Note. x = correct or incorrect response, NA = missing response

A detailed analysis of Reading found a subject with a higher cumulative score than other subjects, but when estimated using theta, it turned out to be lower theta. In fact, based on the Math estimation results, although there are subjects with the same conditions, the difference is not too far. Unlike Reading, for example, Table 5 presents the cumulative score and theta with the response pattern. Cumulative score 10 with NA responses in 9 items (1.1504), getting a higher theta than cumulative score 15, 11, and 16. This is due to more NA responses in Reading than Math. According to preliminary analysis, Reading has 326 NA responses, while Math has only 139 NA responses.

Table 5. Comparison of ability estimation based on response patterns in Reading

CS	Theta	Response Pattern
15	.8898	x x x x x x x x x x x x x x x x x x x x x x
11	.9155	x x x x x x x x x x x x x x x NA NA NA NA NA NA NA
16	1.1428	x x x x x x x x x x x x x x x x x x x x x x
10	1.1504	x x x x x x x x x x x x x NA NA NA NA NA NA NA NA NA

Note. CS = cumulative score, x = correct or incorrect response, NA = missing response

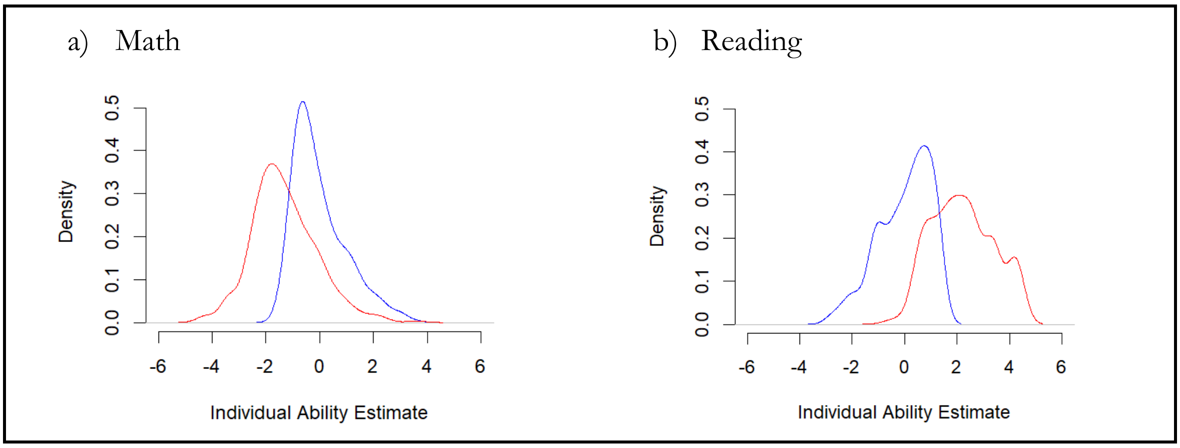
Paired Samples T-test between Cumulative Score and Rasch

Prior to conducting the t-test, z-scores were calculated for cumulative scores in Math and Reading to standardize them to theta. A paired samples t-test was then performed to compare cumulative scores and theta in Math and Reading. The cumulative score was assigned to Group 1 and theta to Group 2. Table 6 displays the t-test results, which were all statistically significant with a p-value < .001. The t-test results for cumulative score and theta in Math show t-values of 75.583 and -58.768, respectively. The size effect generated from Cohen's d in Math and Reading is 4.235 and -3.301, respectively, indicating a very high size effect for both.

Table 6. Paired samples t-test results between cumulative score and theta

Aspect	Cumulative Score (Z-Score)		Theta		*t*(316)	*P*	Cohen's *d*
	M	*SD*	*M*	*SD*			
Math	0	1	-1.3074	1.2093	75.583	< .001	4.235
Reading	0	1	2.2231	1.1758	-58.768	< .001	-3.301

10.21831/pep.v28i1.71661



Note. Blue line = cumulative score, red line = theta

Figure 3. Cumulative score and theta density plots

The t-test results indicate statistically significant differences between cumulative score and Rasch with a 99.9% confidence interval in Math and Reading. Despite that, Math shows a positive t-test and Reading shows a negative t-test. Figure 3 visualizes the comparison of cumulative score and theta data distribution in Math and Reading, highlighting the difference in the t-test results: In Math, the cumulative score data distribution is located to the right of the theta data distribution, resulting in a positive t-test. In contrast, in Reading, the cumulative score data distribution is located to the left of the theta data distribution, resulting in a negative test of difference.

Discussion

The aim of this study is to investigate whether the position of individuals based on ability estimation using cumulative score remains the same or changes when compared to ability estimation using Rasch. The data analysis revealed a difference in individual position based on ability estimation between cumulative score and Rasch. Firstly, the variation of the Rasch score is greater than that of the cumulative score. Secondly, the Rasch score is affected by NA response patterns. Additionally, the paired samples t-test results showed a significant difference between the cumulative score and Rasch score in Math and Reading. The t-test result for Math was positive (75.583), while the t-test result for Reading was negative (-58.768).

The cumulative score has a lower score variation than the Rasch score in both subjects. In Math, the cumulative score variation is 19, whereas the theta score variation is 72. This is because subjects who achieve the same cumulative score can receive different scores when estimated using Rasch. For example, based on Table 2, a subject who has a cumulative score of 3, when his cumulative score is converted to theta, can have 1 of 6 possible theta variations.

Theta variations are caused by different response patterns. The statement is supported by the findings. Specifically, subjects who receive the same cumulative score when estimated with theta (Rasch) are grouped based on NA response patterns and exhibit differences in theta scores. In Math, subjects with a cumulative score of 3 are divided into six variations of theta and grouped based on NA response patterns. Rasch's estimation of ability is influenced by item difficulty and subject response patterns (Baker & Kim, 2017). Furthermore, item difficulty has an impact on subjects' NA responses. For instance, based on Table 2, subject may receive a higher theta score (-1.9794) than another subject (-2.0759) due to the fact that theta -1.9794 has one NA response on items with lower difficulty (-1.0453), whereas theta -2.0759 has one NA response on items with higher difficulty (.8102).

Additional evidence of Rasch accounting for NA response patterns is demonstrated in Reading. The study found that subjects with the same cumulative score, but different numbers of NA responses, had higher theta scores if they had more NA responses. It is important to note that this happens only when subjects are compared with the same cumulative score. For example, in Table 5, a subject with a cumulative score of 10 and NA responses in 9 items (1.1504) had a higher theta score than a subject with a cumulative score of 15 and no NA responses (.8898). When there are more NA responses, the estimated parameters use less information, which decreases the precision of the estimate and increases the standard error (Waterbury, 2019). Additionally, Rasch does not estimate the position of correct or incorrect answers. As long as the cumulative score and NA response pattern are the same, the subject will receive the same theta score. In Math, subjects 16 and 28 correctly answered items M10, M13, M20, M8, M10, M19 respectively, resulting in a theta of -2.0989 for both.

This study utilizes the t-test to measure the difference in cumulative score and Rasch, which distinguishes it from previous studies that have focused on the positive correlation between cumulative score and theta, with correlation results above 0.95 (Fan, 1998; Kiliç, 2019; MacDonald & Paunonen, 2002; Progar & Sočan, 2008). The t-test results for cumulative score and theta in Math indicate a significant difference, with a t-value of 75.583. Similarly, there was a significant difference in the cumulative score and theta for Reading (t=-58.768). A positive t-test result indicates that the cumulative score distribution is to the right of the theta distribution, while a negative t-test result indicates the opposite.

It can be concluded that both cumulative score and Rasch have their own advantages in estimating the subject's ability. Cumulative score estimation is easily understandable by both test makers and test users, as it only counts correct answers. This traditional approach still attracts researchers in test development and analysis due to its theoretical and practical simplicity (Bichi et al., 2019). However, by only presenting the total number of correct answers, the cumulative score does not take into account the weight of each item. As a result, all items are assumed to have the same weight in the cumulative score. Therefore, the cumulative score cannot provide any information other than the total number of correct responses.

In contrast to the cumulative score, Rasch's estimation takes into account not only correct and incorrect responses but also the various NA responses produced by the subject. This allows for more variations in ability estimation than the cumulative score. Kiliç's (2019) research also found that theta estimation had more variation than the cumulative score. Furthermore, Rasch analysis can estimate the probability of a subject answering an item, which is referred to as item difficulty. This is supported by DeMars' (2010) and Kiliç (2019) statement that when using cumulative scores, the item characteristic has no effect on the estimation of individual ability. Therefore, the item characteristic is an advantage of IRT over CTT. The use of Rasch analysis allows for a more comprehensive evaluation of the items necessary for a valid assessment of individual ability and the appropriateness of the items in measuring the intended outcome (Bichi et al., 2019). While Rasch's analysis provides more information, the resulting ability estimates cannot be interpreted directly without making simplifying assumptions. This is because the estimates are still intervals that typically fall within the range of -4 to 4 (Baker & Kim, 2017). To aid interpretation, researchers often provide conversion tables for raw scores, Rasch results, and scale scores (Rasch, 2007).

The presented comparison between cumulative score and Rasch for ability estimation can serve as a reference for test makers when selecting scoring methods. Each method has its own advantages and disadvantages in estimating the subject's ability. If test developers and users prefer an easy-to-apply scoring method, they can use the cumulative score method. However, if they require a scoring method that provides more information, they should use the Rasch method. This way, test makers can determine which method is suitable for the test being constructed.

CONCLUSION

The t-test results demonstrate a significant difference between the two methods. This difference in ability estimation is due to variations in score. Therefore, two individuals with the same cumulative score may receive different Rasch estimates. The study found that Rasch score variation is caused by different NA response patterns and item difficulty. Furthermore, it was discovered that Rasch provides more information than cumulative score as it can estimate the probability of an individual answering an item.

ACKNOWLEDGMENT

This work was supported by the Faculty of Social and Political Science, Brawijaya University.

Conflict of interests

There are no known conflicts of interest associated with this publication.

REFERENCES

Abedalaziz, N., & Leng, C. H. (2018). The Relationship between CTT and IRT approaches in analyzing item characteristics. *MOJES: Malaysian Online Journal of Educational Sciences*, *1*(1), 64–70. http://jice.um.edu.my/index.php/MOJES/article/download/12857/8251

Amelia, R. N., & Kriswantoro, K. (2017). Implementation of Item Response Theory for Analysis of Test Items Quality and Students' Ability in Chemistry. *JKPK (Jurnal Kimia Dan Pendidikan Kimia)*, *2*(1), Article 1. https://doi.org/10.20961/jkpk.v2i1.8512

Anastasi, A., & Urbina, S. (2002). *Psychological testing*. Prentice Hall.

Baker, F. B., & Kim, S. H. (2017). *The Basics of Item Response Theory Using* R. Springer International Publishing. https://doi.org/10.1007/978-3-319-54205-8

Bichi, A. A. (2016). Classical Test Theory: An introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies*, *2*(9), 27–33. https://www.academia.edu/download/53191152/CLASSICAL_TEST_THEORY_An_Introduction_to.pdf

Bichi, A. A., & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education*, *7*(2), 142–151.

Bichi, Talib, R., Atan, A., Ibrahim, H., & Yusof, S. (2019). Validation of a developed university placement test using classical test theory and Rasch measurement approach. *International Journal of Advanced and Applied Sciences*, *6*, 22–29. https://doi.org/10.21833/ijaas.2019.06.004

Champlain, A. F. D. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Columbia University. (2016, August 5). *Rasch Modeling*. Columbia University Mailman School of Public Health. https://www.publichealth.columbia.edu/research/population-health-methods/rasch-modeling

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An empirical comparison of their item/person Statistics. *Educational and Psychological Measurement*, *58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Fernanda, J. W., & Hidayah, N. (2020). Analisis kualitas soal ujian statistika menggunakan Classical Test Theory dan Rasch Model. *Square: Journal of Mathematics and Mathematics Education*, *2*(1), Article 1. https://doi.org/10.21580/square.2020.2.1.5363

Hambleton, R. K., & Jones, R. W. (1993). Comparison of Classical Test Theory and Item Response Theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. SAGE.

Hayat, B., Putra, M. D. K., & Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the Rasch model from three different traditions. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(1), Article 1. https://doi.org/10.21831/pep.v24i1.29871

Kemendikbud. (2019). *Pendidikan Di Indonesia Belajar Dari Hasil PISA 2018*. Pusat Penilaian Pendidikan Balitbang Kemendikbud. https://simpandata.kemdikbud.go.id/index.php/s/tLBwAm6zAGGbofK

Kiliç, A. F. (2019). Can Factor Scores be used instead of Total Score and Ability Estimation? *International Journal of Assessment Tools in Education*, *6*(1), Article 1. https://doi.org/10.21449/ijate.442542

Linacre, J. M. (1994). *Sample Size and Item Calibration or Person Measure Stability*. https://www.rasch.org/rmt/rmt74m.htm

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of Item and Person Statistics based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement*, *62*(6), 921–943. https://doi.org/10.1177/0013164402238082

Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1–11. https://ssrn.com/abstract=1426043

Maulani, M. R., & Rahardjo, B. (2014). Teori pengukuran pendidikan menggunakan Classical Test Theory dan Item Response Theory. *Competitive*, *9*(1), Article 1. https://ejurnal.ulbi.ac.id/index.php/competitive/article/view/257

10.21831/pep.v28i1.71661

OECD. (2017). What is PISA? In *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://read.oecd-ilibrary.org/education/pisa-2015-assessment-and-analytical-framework/what-is-pisa_9789264281820-2-en#page1.

OECD. (2019). *2018 Database—PISA*. https://www.oecd.org/pisa/data/2018database/

Prieto, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes*. 10.1186/1477-7525-1-27

Progar, Š., & Sočan, G. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. *Horizons of Psychology*, *17*(3), 5–24. http://psiholoska-obzorja.si/arhiv_clanki/2008_3/progar_socan.pdf

Rasch. (n.d.). *Rasch Dichotomous Model vs. One-Parameter Logistic Model 1-PL*. https://www.rasch.org/rmt/rmt193h.htm.

Rasch. (2007). *Raw Score-to-Measure (Scaled Score) Tables*. https://www.rasch.org/rmt/rmt211j.htm

Rusch, T., Lowry, P. B., Mair, P., & Treiblmaier, H. (2017). Breaking free from the limitations of Classical Test Theory: Developing and measuring information systems scales using Item Response Theory. *Information & Management*, *54*(2), 189–203. https://doi.org/10.1016/j.im.2016.06.005

Saifuddin, A. (2002). *Tes Prestasi fungsi dan pengembangan pengukuran prestasi belajar*. Pustaka Pelajar Offset.

Sarea, M. S., & Ruslan, R. (2019). Karakteristik butir soal: Classical Test Theory VS Item Respone Theory? *DIDAKTIKA : Jurnal Kependidikan*, *13*(1), 1–16. https://doi.org/10.30863/didaktika.v13i1.296

Vincent, W., & Shanmugam, S. K. S. (2020). The role of classical test theory to determine the quality of classroom teaching test items. *Pedagogia : Jurnal Pendidikan*, *9*(1), Article 1. https://doi.org/10.21070/pedagogia.v9i1.123

Waterbury, G. (2019). Missing Data and the Rasch Model: The Effects of Missing Data Mechanisms on Item Parameter Estimation. *Journal of Applied Measurement*, *20*, 1–12. https://pubmed.ncbi.nlm.nih.gov/31120433/

Widoyoko, E. P. (2012). *Teknik penyusunan instrumen penelitian*. Pustaka Pelajar.

Xu, T., & Stone, C. A. (2012). Using IRT Trait Estimates Versus Summated Scores in Predicting Outcomes. *Educational and Psychological Measurement*, *72*(3), 453–468. https://doi.org/10.1177/0013164411419846