# Item Analysis for Examination Test in the Postgraduate Student's Selection with Classical Test Theory and Rasch Measurement Model

Dedy Triono
Department of Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya 60264, Indonesia
dedytriono@staf.unair.ac.id

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya 60111, Indonesia
riyanarto@if.its.ac.id

Kelly R. Sungkono
Department of Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
kelly@its.ac.id

***Abstract*— University entrance exams are conducted to ensure applicants' qualifications are placed into the program of their choice. Test results have important and significant value in making the right decision about the suitability of the applicant; the validity of the exam is significant to achieve the objectives set. The purpose of this study is to provide empirical evidence of the validity of the new construct in developing the Academic and English Test Exams using the Classical Test Theory and the Rasch Measurement Model. Admission Test for postgraduate entrance examination consisting of 120 multiple choice items with five answers/option choices (A-E) and has been developed and assessed by experts who are competent in their fields and questions are given to 409 postgraduate entrance exam participants. Software applications used for CTT and Rasch Model are ITEMAN version 3 and JMETRIK version 4 windows, where the application is free of licenses. The software automatically generates parameter estimation recommendations for assessing the quality of test items. The CTT results identified 39 questionable items using difficulty and index discrimination. Rasch's results show that the statistics of people (Separation 2.55> 2.00 and reliability 0.87> 0.80) and item statistics (Separation 9.4> 3.0 and reliability 0.99> 0.8) are excellent person and item reliability. Overall, using the Rasch model obtained 68 constructs that incorrectly matched items, as well as irrelevant identified, are suggested to be removed. While CTT provides limited information on two parameters, Rasch's results provide very detailed information about the quality of the items being tested. Thus the two models can be integrated to produce sufficient evidence of validity and reliability items in the development of standardized tests. Even from the second approach, the model produced 28 items in common as problem items. These results indicate that more items are recommended for removal by the Rash model than the CTT can be linked to the procedure followed by two frameworks in determining the quality of test items*.

***Keywords*— C.T.T., Rash Model, Item Analysis**

## I. INTRODUCTION

To get high-quality question instruments, in addition to theoretical analysis (item review), empirical analysis is also necessary. This practical item analysis can be divided into two, namely: with the classical test theory approach and item response theory (IRT) [1] The test is a measurement technique designed as a systematic procedure for studying the behaviour of individuals or groups of individuals [2]. In this description, two analytical methods that are generally used in developing tests, namely traditional or standard item analysis in classical experiments or Classical Test Theory (CTT) and modern interpretation, are based on item response theory (IRT). These processes generally follow the identification of the objectives of the test and the preparation of a pool of items in the test preparation process. To produce tests in educational measurements, the criteria and guidelines that have been established for the development of valid and reliable tests must be followed adequately. This provides accurate information in the use of tests and construction [1]

Analysis of test instruments in education can be done through two approaches. The first approach is the most common and is widely applied in education to date, especially in research, namely the classical test theory (CTT). This statement is following the report [3] in his study entitled "the accuracy of the results of item analysis according to classical test theory and item response theory in terms of sample size," that classical test theory (CTT) is a popular analytical technique. It is used in stock in this century. The conventional test theory developed by Charles Spearman in 1904 can be used to predict the results of an exam. In classical test theory, the aspects that largely determine the quality of the items are the level of difficulty and the distinguishing features of the questions. However, the characteristics of items produced by classical test theories are inconsistent (changing) depending on the ability of test-takers. According to [4], measurement errors in classical test theory can only be sought for groups, not individuals. A second approach is a modern approach with the Rasch model coined by Dr Georg Rasch is a Danish mathematician. Rasch modelling exists to overcome

weaknesses in classical test theory. Rasch modelling provides a different approach to the use of exam scores or raw data in the context of educational assessment. The aim is to produce a measurement scale with the same interval that can later provide accurate information about the test taker as well as the quality of the questions being worked on. In other words, the analysis of the Rasch model will produce information about the characteristics of items and students that have been formed into the same metric [5]. In this study, a comparative analysis of the quality of test instruments will be carried out on the elements of validity, reliability, level of difficulty, and distinguishing features of the questions through the two approaches described above, namely the classical test theory and the Rasch model.

## II. RELATED WORK

### A. Classical Test Theory (CTT)

Classical test theory adopts a deterministic approach (certainty) wherein the main focus of the analysis is the total individual score (X). Each test has an error (E) that accompanies each measurement result in measuring human nature. Pure scores (T) and errors (E) are both latent variables, but the purpose of testing is to conclude individual absolute scores. The score of each item can also be ascertained right and wrong, i.e., for example, if someone's answer is correct, then given a score of 1 and given a score of 0. While IRT focuses on the probability of answering each item where assessing solutions is not on someone's total score but considers one's response/answer at the level of the question. Giving a rating is also not by determining a score of 1 or 0, but the probability of the person getting a score of 1 or a score of 0. The mathematical formula is called CTT Model is represented in Eq. (1)

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

This assumption states that the relationship between visible scores (X), pure scores (T), and measurement errors (E) is additive. The visible score (X) obtained by an individual is an accumulation of absolute ratings (T) and measurement error (E).

### B. Classical Test Item Analysis

In the test preparation process, items that have been qualitatively reviewed by experts in their fields can be declared valid in content. However, in the achievement test, it is necessary to do additional analysis aimed at obtaining items that have a high degree of measurement and power difference so that the goal of size is to distinguish the abilities of one individual from another individual can be achieved. This procedure is often referred to as item analysis and selection because the purpose of this procedure is nothing but knowing which items are feasible to be maintained or revised or even discarded. The process of analyzing and selecting item items based on classical test theory pays attention to three parameters, namely (1) item difficulty level, (2) item discrimination power, and (3) distractor effectiveness [6]. The analysis was carried out based on the subject's answers to the items in the test. Even though the level of difficulty of questions and the discrimination power of things are calculated separately, in the evaluation of both issues, the item is seen as a unitary component that will determine whether an item is considered good or not. The third parameter, namely, the effectiveness of the distractor, only applies to questions in the form of multiple choice.

### C. Level of difficulty

The item difficulty index, as stated by [7], is the "proportion of examinees who get that item correct." The statement explains that the level of difficulty of test items is a number that shows the proportion of test participants who can answer the question correctly. In comparison, the level of difficulty of the test set is a number that shows the average percentage of test participants who can answer all the test sets. The formula used to determine the level of difficulty is as follows Eq. (2)

$$P = \frac{nB}{n} \tag{2}$$

information:
*p*: level of difficulty of the test item
*nB*: number of subjects answering correctly
*n*: total number of subjects

The mathematical model states that the level of difficulty of the problem (*p)* is influenced by the number of participants who worked on the questions correctly divided by the total number of participants present.

As stated by Allen & Yen, the problem of good things is from 0.3 to 0.7. Items with difficulty levels below 0.3 are considered heavy items, whereas if the index is above 0.7, items are deemed secure [7]. Thus the difficulty level (P) criteria can be written in TABLE I.

TABLE I. INTERPRETATION OF ITEM DIFFICULTY INDEX

| Difficulty of Index | Interpretation |
|---|---|
| $P \leq 0.30$ | Difficult |
| $0.31 \leq P \leq 0.70$ | Moderately difficult |
| $P > 0.70$ | Relatively easy |

### D. Item Discrimination Power

The power of discrimination (discrimination) of a test item is the ability of an object to distinguish between high-ability and low-ability test takers [7]. This understanding explains that the power of different test items is the ability of the test items to differentiate between participants in the upper group test (high) and lower group test participants (weak). The formula for calculating the power of different test items is as follows Eq. (3).

$$D_B = \frac{nB_A}{n_A} - \frac{nB_B}{n_B} \tag{3}$$

The formula states that the difference in test items is highly dependent on the results of the whole group of the top answering questions correctly reduced by the entire group of the bottom of the results are positive, then the items can be

accepted. If the "$D_B$" is negative, the problem is terrible and must be discarded. Information:

information:

$nB_A$ : number of subjects who answered correctly in the upper group

$nB_B$ : the number of participants in the lower group who answered correctly

$n_A$ : number of items in the top group

$n_B$ : the number of participants in the smaller group

So these statistics show the extent to which a test successfully distinguishes between people with high ability and people with low knowledge. Different power groupings according to [8], are presented in TABLE II.

TABLE II. DIFFERENTIAL POWER CRITERIA (DB)

| Item Discrimination | Quality of Item |
|---|---|
| D ≥ 0.40 | Good question |
| 0.30 ≤ D ≤ 0.39 | Questions received and corrected |
| 0.20 ≤ D ≤ 0.29 | Problem corrected |
| D ≤ 0.19 | Problem rejected |

Thus, item parameters such as the difficulty index and discrimination index are characteristics that depend on the sample group used to calculate it. If the test group has a high ability, the difficulty index of the test items will below. But on the contrary, if the test group has a low skill, then the index of the difficulty of the test items will be high, likewise, with the characteristics of other test items. So the value of the components of the questions will be influenced by the ability of one group of test-takers.

### *E. Item Reliability (Test level)*

Reliability comes from the word reliability, which can be interpreted as something that can be trusted. In the same case, Drost states that "reliability is a major concern when a psychological test is used to measure some attributes or behaviour" [9]. This definition says that reliability is trustworthiness, reliability, constancy, consistency, or stability. There are several types of safety, namely: (1) internal consistency, (2) security, and (3) equivalent. The internal consistency reliability of the measuring instrument can be calculated using the formula AlphaCronbach, Kuder-Richardson (KR20 or KR21), and the Split Technique. Suparwoto states that the AlphaCronbach coefficient can be used for item analysis with a score of true and false 0, or with a score of 1, 2, 3 sequentially and this method is an effort to determine the reliability coefficient of the instrument/test that refers to the concept of internal consistency [10] The formula used to calculate the Alpha-Cronbach ratio is as follows Eq.(4).

$$r_{1.1} = \left(\frac{k}{k-1}\right)\left(\frac{SD_t^2 - \sum SD_i^2}{SD_t^2}\right) \quad (4)$$

where the reliability coefficient of the test is influenced by the number of items (k) multiplied by the results of the distribution of the score variance of the question. i with the total score variance.

information:

$r_{1.1}$: the reliability coefficient of the test device

$k$: many test items

$SD_i^2$: score variance per item

$SD_t^2$: whole score variant

The reliability level of the instrument can be determined from the value of r can be seen in TABLE III.

TABLE III. INSTRUMENT RELIABILITY LEVEL (R)

| Item Reliability | Level of item |
|---|---|
| r ≤ 0.20 | Very Poor |
| 0.20 < r ≤ 0.40 | Poor |
| 0.40 < r ≤ 0.60 | Medium |
| 0.60< r ≤ 0.80 | High |
| 0.80< r ≤ 1.00 | Very High |

### *F. Effectiveness of the distractor*

Each multiple-choice test has one question and several answer choices. Among the choices of answers, only one is correct. Apart from the right answer, it is the wrong answer. The wrong answer is what is known as a distractor. Thus, the effectiveness of the distractor is how well the wrong choice can fool test-takers who do not see the answer keys available. The more test takers choose the distractor, and the distractor can perform its function correctly. How to analyze the role of the distractor can be done by analyzing the pattern of the dissemination of answer items. The design of the distribution of answers, as said [11], is a pattern that can illustrate how the test taker can determine the choice of solutions to the possible answers that have been paired on each item. According to Depdikbud (1993: 27), a distractor can be said to function correctly if chosen by at least 5% for 4 choices of answers and 3% for 5 options of solutions. Meanwhile, according to Fernandes (1984: 29), distractors are said to be good if chosen by at least 2% of all participants. Distractors who do not meet these criteria should be replaced with other distractors who may be more attractive to test takers to select them.

Deceiver needs to be made in such a way as to attract the attention of test-takers who do not yet have a good concept of the material being tested. [7] states that a minimum good deception indexed 0.1 in the form of a biserial point correlation coefficient, positive value for the answer key, and negative value for a trick.

## III. METHODOLOGY

### *A. Design*

The purpose of this study was to analyze the level of difficulty of items and the ability of people to use two measurement frameworks, namely the Classical Test Theory (CTT) and the Rasch Measurement Model (R.M.M.). A total of 410 undergraduate students from various Departments of the Faculty of Education conducted the 2018 Postgraduate entrance examination, which included the Academic Potential

Test and the English Research Methodology. This consists of 120 multiple choice items with five answers/option (A-E), where is the time for the exam 150 minutes for TPA and English test where time and participants support the stability of data from Classical Test theory.

### B. Data Analysis

Data analysis using Classic Item Analysis and Rasch Model Approach is using the Iteman Software Application version 3 for Classic Analysis and JMetrik version 4 windows for Rasch Analysis. The parameters used to assess item quality in CTT are Item Difficulty, Discrimination, and Reliability. In Rasch analysis, three different stages of estimation are considered, (i) Calibration of test-takers' abilities and item difficulties (ii) Match estimation (iii) Assessment of unidimensionality using Principal Component Analysis (PCA) of Rasch residues [12].

### C. IRT METHOD

In the CTT method, the item difficulty level depends on the ability of the test taker. If the test taker's skill is high, the item difficulty level is low, and vice versa, if the test taker's ability is little, then the item difficulty level becomes high. The level of item discrimination and reliability depends on the heterogeneity and distribution of the test-takers' capabilities. The ability of test-takers is interpreted in the correct number of scores. In IRT, the strength of participants is not affected by the characteristics of the items, and the features of the questions are not affected by the ability of individuals. The essence of IRT is the level of difficulty of objects, and individual skills are measured on the same scale. So that a match is needed between the model and the data. IRT is a statistical theory that contains a mathematical model that states the probability of specific responses to individual items as a function of one's abilities and particular characteristics of an object [13]. The item response theory is also often referred to as the latent trait theory, which is a very significant development in the fields of education and psychology measurement.

The latent trait theory uses three primary concepts in developing measurement models, namely potential space dimensions, local independence, and item characteristic curves [13]. This theory states that a person's behaviour can be explained to a certain degree for the characteristics of that person. These characteristics vary, for example, verbal ability, quantitative, psychomotor. This characteristic is also called a trait. A person's position on a character can be used to estimate the magnitude of the person's abilities. This trait is often expressed as a person's ability dimension. The three logistical parameter model (3PL) is parameter a (different power), parameter b (difficulty level), parameter c (guess) when a person's possible correct response to a particular item is expressed as one's ability. Furthermore, this expression is referred to as the Item Characteristics Curve (ICC). The two-parameter logistics model (2PL) is parameter a, parameter b, and parameter c. It is assumed that everyone who has low testability has no chance of success in answering the item (c = 0). The one logistic parameter model (1PL) or known by the name of the Rasch model, is parameter b, parameter a is assumed to be equal to 1. In contrast, parameter c is considered to be zero (c = 0). Estimation of a person's abilities and estimation of item parameters from a model is selected and obtained from data provided by respondents (test-takers).

## IV. GUIDELINES FOR ITEM ANALYSIS USING THE RASCH MODEL

### A. One Parameter Logistic (Rasch Measurement Model)

There are three widely used IRT models namely One-Parameter Logistics Model (1-PL), Two-Parameter Logistics Model (2-PL) and Three-Parameter Logistics Model (3-PL) where each of these models has their parameters One key component that distinguishes this model is the Item Characteristics Curve (ICC) which graphically displays information of each item produced by I.R.T. One-Parameter Logistics Model (1-PL) also known as Rasch Model is the most basic model in IRT which estimates only one parameter, the difficulty parameter (b) [14]

In 1- PL, item discrimination level (a) and guess probability (c) is assumed to be constant [15]. In the 1-PL model, the ICC for each item is given the equation below Eq. (5).

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (5)$$

where $P_i(\Theta)$ is the probability of students with ($\Theta$) ability to respond to item-i correctly, and b is the level of difficulty of item-i. The bi value typically ranges from -2 to 2 but can take more extreme values. As noted in [14], b and scaled using a normal distribution with a standard deviation of 1.0 and a mean of 0.0, hence [15] presents two summaries of this equation:

(i) The smoother the item is, the high probability students will answer it correctly

(ii) Students with high ability are more likely to answer questions correctly compared to students with low ability.

### B. Item Compatibility Level

Matching items mean that they behave consistently with what is expected by the model. If it is found that the questions are not fit, this is an indication that there is a misconception in students about the item. The fit index provided in the Rasch analysis is ZSTD Person Infit, ZSTD Person Outfit, MNSQ. Infit Person, MNSQ. Person Outfit, ZSTD Infit Item, ZSTD Outfit Item, MNSQ. Infit Item, MNSQ. Outfit Item [16].

MNSQ. Values are always positive and move from zero (0) to infinity ($\infty$). In this case, the MNSQ. Value is used to monitor the suitability of the data with the model. The expected mean square value is 1 (one). A mean-square value for infit or outfit higher than one, say 1.3, indicates that the observed data has 30% more variation than predicted by Rasch. The infit or outfit value is less than 1, say 0.78 (1-0.22 = 0.78), indicating that the observed data has 22% fewer variations than predicted by the Rasch model [12]. At the same time, the expected value of z is close to 0 (zero). When

the data observed is following the model, the amount of z has an average approaching 0, and the standard deviation is 1. The ZSTD value that is too large (z> +2) or too low (z <-2) indicates the items do not match the expected model. Standardized z values (ZSTD) on infit and outfit can be either positive or negative. A negative ZSTD value indicates less variation compared to the model. Response answers approach the Guttman-style response string model that all subjects with high ability can answer correctly, and all questions with low ability answer incorrectly on these items. At the same time, positive values indicate that the variation of solutions is more than in the model. Response responses are irregular and unpredictable [12].

According to [16], the criteria used to check the appropriate item are :

1. The appropriate Outfit Mean Square (MNSQ) value: 0.5 <MNSQ <1.5
2. The appropriate Z-standard (ZSTD) outfit values: -2.0 <ZSTD <+2.0

If the items in the two criteria are not fulfilled, it means the items are not good and need to be revised or replaced. Unlike the level of difficulty of consistent items, the level of suitability of this item is strongly influenced by the size of the sample size. The answer key error, caused by the large number of participants working on careless problems, and questions that have low discrimination so that it can reduce the suitability value of items. Another thing to be noticed is, this ZSTD value is susceptible to the number of samples. If the sample used is large (> 500), there is a tendency for this ZSTD value to have a value above 3. Therefore, some experts recommend not using this ZSTD criterion if the sample used is large enough [17]

### *C. Rasch Discrimination Power (Point Measure Correlation)*

The Rasch Discrimination Power or item score and Rasch (Pt Measure Corr) score correlation principle is, in principle, the same as the item discrimination power measured by the CTT approach. It's just that in classical test theory, the computation uses raw scores. The Pt Measure Corr used is measure scores. Pt Measure Corr value of 1.0 indicates that all examinees with low ability answer the items incorrectly, and all test participants with high ability answer the items correctly. While the Pt Measure Corr value is negative, indicating items that are misleading because the examinees with low ability can answer the items correctly and test participants with high ability answer incorrectly. Problems with negative correlation values must be checked to see whether the answer key is wrong, needs to be revised, or deleted from the test [1].

As with classical test theory, the ideal correlation score for grain and Rasch scores is positive and not close to zero. Some experts have opinions about how much Pt Measure Corr is required. Alagumalai, Curtis, & Hungi (2005) classify these values to be very good (> 0.40), good (0.30-0.39), sufficient (0.20-0.29), unable to discriminate (0 - 0.19), and requires examination of items (<0)

### *D. Item Difficulty Level (Item Measure)*

The level of difficulty of the items in the IRT model is the same as the CTT, which is the ratio of the number of correct answers to the number of questions tested. The difference is that the probability value is scaled by entering the logarithmic function. The logarithm estimation results from odd-ratio are called measure values. If in classical test theory, a high difficulty index value means that the problem is easy, in Rasch, the top logit value model indicates the item is delicate. Just like in classical test theory, there is no standard of what level of difficulty is received in the test. This depends on the purpose of the test itself.

## V. RESULT AND DISCUSSION

### *A. CTT analysis result*

The results of item analysis using CTT consider three (3) parameters in judging the quality of items to be used in assessing students' abilities; these are item difficulties (p), Item discrimination (D), and reliability (r). The results are presented in TABLE IV and TABLE V.

Summary statistics presented in TABLE IV shows that, for the total number of 120 items with 409 examinees, the mean score was 68.03 and Standart deviation = 12.52. The mean item difficulty and biserial are 0.57 and 0.35, respectively. These statistics revealed that the test has a sufficient reliability index according to CTT because an index of 0.87, which is higher than the recommended value of 0.70 [1].

TABLE IV. SUMMARY ITEM STATISTICS

| Parameter | Value |
|---|---|
| Total Test Questions | 120 |
| Number of participants | 409 |
| Alpha Reliability Coefficient | 0.870 |
| Average Participant Score | 68.034 |
| Standard Deviation | 12.519 |
| Item Difficulty Level | 0.567 |
| Biserial Average | 0.347 |

The item difficulty averages 0.567 is within the standard required for items that are quite difficult, with a discrimination index of 0.347 without revision for all tests [1]. The results presented in TABLE V below show that CTT Item analysis shows that 81 or 67.5% items have satisfactory item statistics (D> 0.19).

These items are satisfied with the minimum requirements for inclusion into the final version of a test with a minor revision. However, 39 (32.5%) based on the established criteria are recommended to be eliminated from the analysis having (D ≤ 0.19). This means that 39 of these defective items are not appropriate and may not be included in the final draft test. The internal consistency reliability of the test items was assessed and found to be acceptable with Cronbach's alpha value of 0.870 (TABLE IV).

TABLE V. CTT ITEM ANALYSIS CHART

| Difficulty Index | High Difficult (<0.30) | Moderate (0.31≤0.70) | Easy (>0.70) | Total |
|---|---|---|---|---|
| **Excellent ≥ 0.40** | 74,93 | 25,28,57,63,64,66,71,72,78,81,84,88,92,95,96,104,114 | 54,75,80,82 | 23 |
| **Good 0.30 ≤ D ≤ 0.39** | 49,67,97,105,107,113 | 20,27,38,46,48,53,59,60,79,89,101,109 | 1,22,23,24,30,32,33,35,40,43,50,76,86,90,108,65,87 | 35 |
| **Marginal 0.20 ≤ D ≤ 0.29** | 69,73,103,116 | 17,55,61,62,77,91,98,110,111,117 | 15,16,21,26,29,31,44,56,58 | 23 |
| **Poor ≤ 0.19** | 3,4,6,42,45,47,68,83,99,106,115,118,119,120 | 2,5,7,8,9,10,34,36,41,52,70,85,94,100,102,112 | 11,12,13,18,19,37,39,51 | 39 |

## B. *Rasch measurement results (Point Measure Correlation)*

Point Measure Correlation value received: 0.4 <pt measure corr <0.85. Because the point measure correlation is in principle the same as the point-biserial correlation in classical test theory,[1] classify the value of Point Measure Correlation to be very good (> 0.40), good (0.30-0.39), sufficient (0.20-0.29), unable to discriminate (0-0.19), and requires examination of items (<0).

TABLE VI. RASCH ITEM ANALYSIS CHART

| Difficulty Index | High Difficult (<0.30) | Moderate (0.31≤0.70) | Easy (>0.70) | Total |
|---|---|---|---|---|
| **Excellent ≥ 0.40** | 74 | 25,57,66,84,88,104,114 | 54,75,80,90 | 12 |
| **Good 0.30 ≤ D ≤ 0.39** | 93,105,115 | 20,27,28,38,48,63,64,72,78,89,95,96,109 | 23,24,32,33,40,43,65,82,86,108,87 | 27 |
| **Marginal 0.20 ≤ D ≤ 0.29** | 49,67,107,113,118 | 17,46,53,59,60,62,77,79,101,110 | 16,21,22,26,29,30,31,50,56,58,76, | 26 |
| **Poor ≤ 0.19** | 3,4,6,42,45,47,69,73,97,99,103,116,118,119,68,83,106,120 | 52,55,2,5,7,8,9,10,34,36,41,61,85,91,98,111,112,117,70,94,100,102 | 1,11,15,18,35,39,44,51,12,13,14,19,37 | 55 |

## C. *Item Fit Level*

According to [16], the value of means-square outfit, z-standard outfit, and point measure correlation are the criteria used to see the level of conformity of items. If there are items that do not meet the criteria, then the item should be repaired or replaced. Guidelines for assessing item conformity criteria according to Boone et al. (2014) are as follows.

- Accepted Outfit Mean Square (MNSQ) value: 0.5 <MNSQ <1.5
- Received Z-standard (ZSTD) outfit values: -2.0 <ZSTD <+2.0

TABLE VII. ITEM STATISTICS VALUE OUTFIT MEAN SQUARE (MNSQ)

| ZSTD value | Item | Total |
|---|---|---|
| **0.5< MNSQ < 1.5 Received** | All items except items Fail | 92 |
| **0.5 < MNSQ < 1.5 Failed** | 2,3,4,5,6,7,8,9,10,34,36,37,41,45,52,70,83,94,99,100,102,106,111,112,115,116,118,119 | 28 |

TABLE VIII. ITEM STATISTICS Z-STANDARD (ZSTD) OUTFIT VALUE

| ZSTD value | Item | Total |
|---|---|---|
| **-2.0 < ZSTD < +2.0 Received** | All items except items Fail | 85 |
| **-2.0 < ZSTD < +2.0 Failed** | 2,5,6,7,8,9,10,25,27,34,36,41,45,52,54,57,66,70,71,72,74,75,78,81,84,88,92,94,95,100,102104,111,114,115 | 35 |

## D. *Misfitting or problematic items by CTT and Rasch*

The misfitting items, otherwise known as problematic or defective items, were identified using the two approaches. The 'problematic' issues identified by each framework and the everyday items identified are presented in TABLE IX.

TABLE IX. PROBLEMATIC ITEMS DETECTED BY CTT AND RASCH

| Model | Number Detected | Item Deleted |
|---|---|---|
| **Rasch** | 68 | 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,18,19,25,27,34,35,36,37,39,41,42,44,45,47,51,52,55,61,68,69,70,71,72,73,74,75,78,81,83,84,85,88,91,92,94,95,97,98,99,100,102,103,104,106,111,112,114,115,116,117,118,119,120 |
| **CCT** | 39 | 2,3,4,5,6,7,8,9,10,11,12,13,18,19,34,36,37,39,41,42,45,47,51,52,68,70,83,85,94,99,100,102,106,112,115,118,119,120 |
| **Common Item Detected** | 28 | 1,14,15,25,27,35,44,55,61,69,71,72,73,74,75,78,81,84,88,91,92,95,97,98,102,103,104,111,114,116,117 |

After being analyzed based on two methods in which the classical theory of the model produced 39 items with poor discrimination index category (≤ 0.19) while using the Rash model theory, 68 items did not meet the criteria. Even from the second approach, the model produced 28 things in common as problem items. These results indicate that more details are recommended to be deleted by Rash than CCT can be linked to the procedure followed by two frameworks in determining the quality of test items.

Whereas CTT relies on two parameters namely item difficulty and discrimination, Rasch is not limited to item parameters by adding Discrimination Power and Item Fit all contribute to the valuation of things that are not appropriate, Example item 93 was identified by Rasch and CTT as delicate items, but CTT classifies it as Excellent items because the discrimination index ignores the level of difficulty.

Looking at the results, some items that were identified as items that were not appropriate by the CTT were classified as necessary by providing more detailed information based on the ability of the participants. While a participant's expertise in CTT is determined based on a raw (total) score on the test, Rasch's interpretation of the participant's ability is based on

the participant's response to delicate and natural items. In CTT, students with the same total score will be interpreted to have the same capacity.

However, in IRT students with the same total score will be interpreted to have different abilities, if one scores more on a more natural item and another score on a delicate question. Students who print more difficult questions will be interpreted to have higher abilities. Whereas the CTT difficulty score of the item indicates how difficult or easy the subject is in the test for the group of examinees, the Rasch measurement provides a better interpretation of the spread of the item difficulty concerning the test participant's level of ability. Rasch made this feasible through mapping facilities [18]

## VI. CONCLUSION

The main objective of this study is to provide empirical evidence of the validity of the construct as well as the reliability of the Student Entrance Examination Test developed for State Universities using the traditional Classical Test Theory and the Rasch Measurement Model (R.M.M.). More important is to identify the suitability / inappropriate or good or bad items that will be maintained or eliminated from the test when the two CTT and RMM frameworks are used and then identify the strengths and or weaknesses of each of the two approaches in test development and validation.

The findings of this study indicate that more items recommended for removal by Rash than CTT might be related to the technique followed by two approaches in determining the features of test items. Whereas CTT depends on two parameters of item difficulty and discrimination, Rasch is not limited to item parameters; besides item parameters, reliability of people, item maps, fit statistics, and bullies all contribute to the assessment of item incompatibility. Based on these findings, the selection of psychometric procedures depends on many elements. Still, the interpretation using Rasch provides more detailed information about the structure of the items needed for a valid assessment of the ability and suitability of students of the things to measure the desired results.

## ACKNOWLEDGMENT

The authors would like to sincerely thank Institut Teknologi Sepuluh Nopember, the Directorate of Higher Education, Indonesian Ministry of Education and Culture, and LPDP through RISPRO Invitation Program for funding the research.

## REFERENCES

1. Ado Abdu Bichi, R.T., Noor Azean Atan, Halijah Ibrahim, Sanitah Mohd Yusof, *Validation of a developed university placement test using classical test theory and Rasch measurement approach.* International Journal Of Advanced And Applied Sciences, 2019. 6(6): p. 22-29.
2. Ado Abdu Bichi, R.T., Rahimah Embong, Hasnah Binti Mohamed, Mohd Sani Ismail, Abdallah Ibrahim, *Rasch-Based Objective Standard Setting for University Placement Test.* Eurasian Journal of Educational Research, 2019. 19(84): p. 1-14.
3. Hidayati, K., *Keakuratan Hasil Analisis Butir Menurut Teori Tes Klasik dan Teori Respons Butir Ditinjau dari Ukuran Sampel.* 2002.
4. Wahyuni, K.M.d.S., *Analisis Kemampuan Peserta Didik Dengan Model Rasch* in *Seminar Nasional Evaluasi Pendidikan*. 2014 Semarang. p. 9.
5. Waller, S.P.R.a.N.G., *Item Response Theory, and Clinical Measurement.* The Annual Review of Clinical Psychology 2009. : p. 27–48.
6. Petrillo, J., Cano, S. J., McLeod, L. D., Coon, C. D., *Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: a comparison of worked examples.* Value Health, 2015. 18(1): p. 25-34.
7. Afraa Musa, S.S., Abdelmoniem Elmardi, Ammar Ahmed, *Item difficulty & item discrimination as quality indicators of physiology MCQ examinations at the Faculty of Medicine Khartoum University.* Khartoum Medical Journal, 2018. Vol. 11: p. 1477 - 1486.
8. Risa Syukrinda, W.R., *Scoring on Multiple-Choice Test and Achievement Motivation on Geography Learning Outcomes.* American Journal of Educational Research, 2016. Vol. 4 No. 15.
9. Drost, E., *Validity, and Reliability in Social Science Research.* Education Research and Perspectives, 2011. 38: p. 105-124.
10. Setyawarno, D., *Penggunaan AplikasiSoftware Iteman (Item and Test Analysis) untuk Analisis Butir Soal Pilihan Ganda Berdasarkan Teori Tes Klasik.* Ilmu Fisika dan Pembelajarannya, 2017. 1.
11. Jinnie Shin, Q.G.a.M.J.G., *Multiple-Choice Item Distractor Development Using Topic Modeling Approaches.* Frontiers in Psychology, 2019. 10.
12. Andri Syawaludin, Y.S.a.W.R., *RASCH Model Application for Validation of Measurement Instruments of Student Nationalism.* International Conference on Education, 2019. Vol. 5(Issue 2): p. 26-42.
13. Mahmud, J., *Item response theory: A basic concept.* Academic journals, 2017. 12(5): p. 258-266.
14. Tobore, M.A.-i., & Prof. Andrew Igho Joe, *Development and Standardization of Adolescents' Social Anxiety Scale Using the One-Parameter Logistic Model of Item Response Theory.* International Journal of Innovative Social & Science Education Research, 2018. 6: p. 70-76.
15. Magno, C., *Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data.* The International Journal of Educational and Psychological Assessment, 2009. 1(1): p. 1-11.
16. William J.Boone , J.R.S., Melissa S.Yale, *Rasch Analysis in the Human Sciences*. 2014, New York London: Springer Dordrecht Heidelberg.
17. Widhiarso, B.S.W., *Aplikasi Model Rasch Untuk Penelitian Ilmu Sosial*. 2014.
18. Adibah Binti Abd Latif, NFMA, Wilfredo Herrera Libunao,Ibnatul Jalilah Yusof and Siti Sarah Yusri *Multiple-choice items analysis using classical test theory and rasch measurement model.* Man in India, 2016. 96: p. 173-181.