Jurnal Penelitian dan Evaluasi Pendidikan
Volume 27, No 2, December 2023 (133-145)
Online: http://journal.uny.ac.id/index.php/jpep

Comparing item-total correlation and item-theta correlation in test item selection: A simulation and empirical study

Sukaesi Marianti*; Ana Rufaida; Nur Hasanah; Sofia Nuryanti
Universitas Brawijaya, Indonesia
*Corresponding Author. E-mail: s.marianti@ub.ac.id

ARTICLE INFO

Article History
Submitted:
03 June 2023
Revised:
27 June 2023
Accepted:
19 September 2023

Keywords
Item-total correlation; Item-theta correlation; Item selection; Simulation study

Scan Me:

ABSTRACT

One of the important processes in the evaluation of the psychometric properties of a test is item selection. The item selection process usually uses a very popular technique called item-total correlation. This study attempts to describe the item-total correlation technique and explore it using a similar technique called item-theta correlation. Both techniques are applied using simulation studies by creating several conditions related to test length and sample size. After the simulation study, the next step is the study using empirical data as an illustration of the results of the simulation study. The results of this study show that there are differences in the results of item selection based on these two approaches. Item-theta correlation detects more items that have weak discrimination power than item-total correlation. The difference is more noticeable in conditions where the cutoff point used for item selection is low(.20).

To cite this article (in APA style):
Marianti, S., Rufaida, A., Hasanah, N., & Nuryanti, S. (2023). Comparing Item-Total Correlation and Item-Theta Correlation in Test Item Selection: A Simulation and Empirical Study. *Jurnal Penelitian dan Evaluasi Pendidikan, 27*(2), 133-145. doi:https://doi.org/10.21831/pep.v27i2.61477

INTRODUCTION

Item analysis is an important process in the evaluation of the psychometric properties of a test. Test developers often use item analysis as a basis for determining whether items can be selected as good items or need to be revised. The process of item analysis itself involves statistical calculations and is a quantitative process. These calculations can then serve as a basis for eliminating items.

Psychometrics often uses a classical paradigm called Classical Test Theory (CTT). Among the various item selection techniques in the CTT paradigm, item-total correlation is the most frequently used technique (Xie & Cobb, 2020). As its name implies, item-total correlation uses statistical correlation techniques. The two correlated variables are the item and the total score, where the total score is the accumulated score of the entire item in the test. The context of the item-total correlation is a homogeneous test, where the items in the test are those items that measure the same latent construct. In the context of interpreting the item-total correlation value, a high correlation value indicates homogeneity. Conversely, a low correlation value indicates that the test has low homogeneity.

Furthermore, the item-total correlation also indicates the discrimination power of an item, which indicates the ability of an item to distinguish subjects who have high, medium, or low ability (Desjardins & Bulut Boca Raton, 2019; Macdonald et al., 2002; Wu et al., 2016). The higher the correlation value, the more powerful an item is in discriminating between the abilities

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

10.21831/pep.v27i2.21477

of test takers (Andrich & Marais, 2019). The high and low levels of discrimination power are also often combined with the use of internal consistency analysis (MacDonald & Paunonen, 2002; Xie & Cobb, 2020).

Researchers are concerned that the total score in the CTT approach, obtained through the total score of each item, may not accurately reflect the position of the test taker because it is on an ordinal scale and cannot accurately estimate the test taker's true score (Xie & Cobb, 2020). ven so, there is something that concerns the researchers: the total score in the CTT approach is considered rough because it is on an ordinal scale and cannot accurately reflect the position of the test taker. Researchers can explore item-total correlation in the CTT paradigm by replacing the total score with theta, an ability parameter in the Item Response Theory (IRT) paradigm. Theta is an ability parameter in the Item Response Theory (IRT) paradigm. The ability of the test taker is described by theta, and it is estimated on the interval scale. The theta estimation process involves the pattern of test takers' responses to a set of items. Bock and Gibbons (2021) posited that a critical difference between IRT-based estimates of theta and the total score derived from CTT is the precision of measurement. In longitudinal studies, the estimated theta may exhibit small and statistically insignificant changes over time. While the total score in the CTT paradigm can be easily obtained by accumulating the score of each item, However, the total score value is on an ordinal scale without considering the test taker's response pattern to a set of items. This leads to many similar total scores in a score distribution (Hu et al., 2021; MacDonald & Paunonen, 2002).

In this study, we utilize dichotomous data, which consists of 1 representing correct and 0 signifying incorrect. This type of data is typically obtained from cognitive tests, scored based on the accuracy of responses. The dichotomous data in this research is derived from simulated data generated by the IRT model. In addition, we also utilize real-world data from the Trends in International Mathematics and Science Study (TIMMS). TIMMS is an international evaluation conducted every four years to assess the educational achievements of students in mathematics and science. It gauges the knowledge and skills of students in these subjects, yielding invaluable data on the performance of educational systems globally (Yilmaz & Keskin, 2020).

Based on the advantages and disadvantages of each paradigm, this study will apply the CTT paradigm for item selection by correlating items with totals and correlating items with theta. Let x be an item, then in the first technique (item-total correlation), y is the total score. In the second technique (item-theta correlation), y is theta. The results of the two techniques will be compared to see if there is a difference in the result of the item selection when the variable Y changes. The comparison will be carried out through a simulation study by creating several test conditions related to the test length and sample size. As an illustration, empirical data will be analysed using the two techniques mentioned above after the simulation study.

RESEARCH METHOD

Item Total Correlation

Item-total correlation is the correlation of an item score to the total score on a test. This correlation is used to test the homogeneity of items or, in other words, to see how uniform the items in a test are in measuring a single concept. Item-total correlation interpretation relies on the magnitude of the correlation coefficient. A higher value shows a more substantial discriminative power of the item (Karakaya & Kiliç, 2021). A high correlation value for tests assessing cognitive skills suggests that an item effectively differentiates people based on their abilities. Bandalos and Deborah, Finch and French (2018), and Supratiknya (2014) mentioned that item-total correlation indicates how well the items measure constructs. Later, the correlation will determine whether the item needs to be eliminated or selected.

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

10.21831/pep.v27i2.21477

There are numerous methods for determining the item-total correlation of an item. One of them is to use point-biserial correlation when the data is a dichotomous response. Variable X in this study refers to an item containing dichotomous data, and Y refers to the overall score calculated by summing all the items' scores. Finch and French (2018), Jacobs and Viechtbauer (2017) presents the formula for the point-biserial correlation as follows:

$$\rho_{pbis} = \frac{\bar{x}_+ - \bar{x}}{s}\sqrt{\frac{P^*}{Q^*}} \quad (1)$$

Where $\bar{x}_+$ is the mean test score for examinees answering item correctly, $\bar{x}$ is the mean test score for all examinees, s is the standard deviation for test scores of all examinees, P^* is the proportion of examinees answering an item correctly, and Q^* is the proportion of examinees answering an item incorrectly.

Upon obtaining the point-biserial correlation coefficient (ρ_{pbis}), the subsequent analysis involves calculating the corrected item-total correlation. This is achieved through the implementation of a formula, as follows:

$$\rho_{i(y-i)} = \frac{\rho_{pbis}\sigma_y - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_y^2 - 2\sigma_i\sigma_y\rho_{pbis}}} \quad (2)$$

Where ρ_{pbis} is the item-total correlation, σ_y is the standard deviation of total scores, and σ_i is the standard deviation of an item.

Item-theta correlation

The item-theta correlation indicates the relationship between an item and the latent trait. This correlation value reveals how effectively an item discriminates persons at various latent trait levels. A high item-theta correlation signifies a strong relationship with the latent trait, effectively discriminating persons possessing different latent traits. Conversely, a low item-theta correlation suggests that the item does not measure the same construct as other items in the test, thus limiting its ability to distinguish individuals at different levels of the latent trait (Guo et al., 2022).

The item-theta correlation is the correlation between a specific item and the latent trait or ability, as estimated based on the Item Response Theory (IRT) models. In this context, theta is treated as an observable variable (Guo et al., 2022). In terms of basic concepts and purpose, the item-theta correlation shares some similarities with the item-total correlation. These two correlations use the same correlation technique, which is point-biserial correlation. However, the variable Y in the item-theta correlation differs from the variable Y in the item-total correlation. If the variable Y in the item-total correlation is the total score, then the variable Y in the item-theta correlation is the theta estimate.

Therefore, before correlating an item and a theta, the theta value must first be estimated using the IRT model. The equation for estimating the level of the latent trait (i.e., ability), as presented by (Baker, 2001), can be expressed as follows:

$$\theta_{s+1} = \theta_s + \frac{\sum_{i=1}^{N} a_i[u_i - P_i(\theta_s)]}{\sum_{i=1}^{N} a_i^2 P_i(\theta_s)Q_i(\theta_s)} \quad (3)$$

Where θ_s is the estimated ability of the examinee within iteration s, θ_s on the right side of the equal sign is set to some arbitrary value, such as 1. The notation a_i is the discrimination

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

10.21831/pep.v27i2.21477

parameter of item i, u_i is the response made by the examinee to item i (0, 1), $P_i(\theta_s)$ is the probability of correct response to item i, and $Q_i(\theta_s)$ is the probability of incorrect response to item i.

Subject

The research utilizes real data from the TIMSS 2019 cognitive test, sourced from the IEA TIMSS & PIRLS International Study Center. The Trends in International Mathematics and Science Study (TIMSS) is a crucial test that assesses the performance of international students in grades 4 and 8 in mathematics and science. The main objective of TIMSS is to enhance future teaching and learning in these subjects. The study analyzed the TIMSS data with a test code of BSAMYSZ7 for the TIMSS Electronic Mathematics test administered in Malaysia in 2019. It consisted of 135 items and involved 534 students from 13 schools in Malaysia. The data collected is dichotomous, where 0 represents an incorrect answer and 1 represents a correct answer.

The research utilizes real data from the TIMSS 2019 cognitive test, sourced from the IEA TIMSS & PIRLS International Study Center. The Trends in International Mathematics and Science Study (TIMSS) is a crucial test that assesses the performance of international students in grades 4 and 8 in mathematics and science. The main objective of TIMSS is to enhance future teaching and learning in these subjects. The study analyzed the TIMSS data with a test code of BSAMYSZ7 for the TIMSS Electronic Mathematics test administered in Malaysia in 2019. It consisted of 135 items and involved 534 students from 13 schools in Malaysia. The data collected is dichotomous, where 0 represents an incorrect answer and 1 represents a correct answer.

Method

The method employed in this study is a psychometric quantitative methodology utilizing the Classical Test Theory (CTT) paradigm. The item analysis was carried out using the item-total correlation approach, and further improved through the item-theta correlation technique. Both correlations have a similar underlying principle but differ in the Y variable used - the total score and theta. The results of the analysis will be used to select items, and this study will compare the outcome of item selection based on the item-total and item-theta correlation analysis.

The research conducts a simulation study using the Markov Chain Monte Carlo (MCMC) method (Bulut & Sünbül, 2017; Feinberg & Rubright, 2016). Some conditions are created to run the simulation study, such as conditions based on test length, sample size, and the percentage of poor-quality items. The simulation study has several steps as depicted in Figure 1.

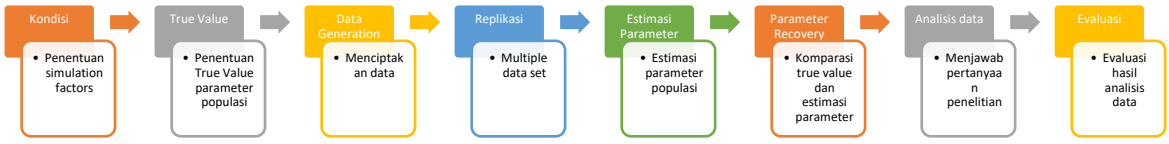

Figure 1. Steps of Simulation Study using Markov Chain Monte Carlo (MCMC) Method

The simulation study will be followed by a real data analysis to reinforce its results. A single dataset will be analyzed using the same steps outlined in the simulation study.

Statsitical Analysis Technique

The data analysis in this research is conducted through a multi-stage process utilizing the CTT paradigm. The primary aim of this analysis is to estimate the correlation between individual items and the total test score, as well as the correlation between items and theta.

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

10.21831/pep.v27i2.21477

The analysis is facilitated by the R programming language, which provides a robust platform for conducting statistical computations (Paek et al., 2019). The package employed in this study is the Test Analysis Modules (TAM) (Robitzsch et al., 2022), capable of executing data analysis based on the Item Response Theory (IRT) model and the mixture model. A more in-depth explanation of the data analysis procedure will be provided in the following sections.

Item Response Theory

According to Luecht and Hambleton (2021), Baker and Kim (2017), de Ayala (2022), three different models, which are based on logistic parameters, can be utilized in Item Response Theory (IRT) approaches, including 1-PL, 2-PL, 3-PL. The choice of logistic parameter model will be tailored based on the number of samples and items present in the data under examination. In the present study, item analysis was performed utilizing the 2-PL model, which encompasses two item parameters: item discrimination power and item difficulty level.

Ability Estimation

Following the estimation of item parameters using the 2PL model, the next step involves the estimation of the examinee's abilities. Subsequently, the estimated abilities are utilized as theta scores in the item-theta correlation analysis.

Correlation Analysis

The correlation between item-total and item-theta is estimated using the point biserial correlation technique, a specialized method for estimating the relationship between two variables X and Y, where variable X is a dichotomous variable. In this study, variable X is an item with a dichotomous response model, and variable Y is the total score or theta value.

FINDING AND DISCUSSION

Findings

This study employed a simulation study utilizing generated data based on various conditions of sample size and test length. Following the completion of data simulation, real data was utilized to illustrate the analysis conducted using the previously generated data. All data analysis was performed using the R programming language.

Simulation

This study utilized a simulation-based approach to generate data under different conditions, including sample size (N=500 and N=1000), test length (K=20 and K=40), and the proportion of low-quality items in the data (20% and 30%). The simulation was conducted through 5000 iterations, with a 500 burn-in iteration process and 50 replications. The results of the parameter recovery analysis are presented in Table 1, while the false alarm rate results are presented in Table 2. The use of R programming language was crucial in conducting this comprehensive and rigorous simulation study.

The dataset comprises of low-quality items was generated through the utilization of two extreme parameters, denoted as a and b. The values of a were generated within a range of -1 to .94, while the values of b were generated within a range of -9.11 to 3.81. Furthermore, the dataset contains high-quality items (with mean ability, mean difficulty, and mean discrimination equal to 0, 0, and 1 respectively) were substituted with low-quality items at two different ratios of 20% and 30%. The contaminated data was analyzed utilizing the r_{it} and $r_{i\theta}$ techniques to

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

10.21831/pep.v27i2.21477

examine their effectiveness in eliminating items based on correlation values that fall below a specified cut-off of .20 and .30.

Table 1. Simulated and estimated parameter values (over 50 replications) 2PL IRT model for 1000 test takers and with K=20 and K=40 items.

Component	True Mean	K=20		K=40	
		Mean	SD	Mean	SD
		N=500			
Person					
Ability	0	0	.622	0	.659
Item					
Discrimination	1	1.067	.121	1.083	.114
Difficulty	0	-.487	.074	-.469	.071
		N=1000			
Person					
Ability	0	0	.621	0	.663
Item					
Discrimination	1	1.005	.093	1.038	.086
Difficulty	0	-.431	.057	-.355	.053

Note. N: the number of test takers. K: the number of items.

Table 1 presents the results of a parameter recovery analysis, demonstrating the true mean, estimated mean, and estimated standard deviation of various conditions under two different sample sizes (N=500 and N=1000), and two different test lengths (K=20 and K=40). In general, the results indicate successful recovery of the parameters, as evidenced by the small discrepancy between the estimated mean and true mean. The true mean represents the actual average value of the data. For N=500, the mean of the ability component is 0 for both test lengths (K=20 and K=40), which is equivalent to the true mean of 0. The mean of the discrimination component is 1.067 for K=20 and 1.083 for K=40, close to the true mean of 1. The mean of the difficulty component is -.487 for K=20 and -.469 for K=40, deviating from the true mean of 0. For N=1000, the mean of the ability component is 0 for both test lengths, equivalent to the true mean of 0. The mean of the discrimination component is 1.005 for K=20 and 1.038 for K=40, close to the true mean of 1. The mean of the difficulty component is -.431 for K=20 and -.355 for K=40, deviating from the true mean of 0.

Table 2. Proportion of eliminated items based on r_{it} and $r_{i\theta}$ using cut-off point .20

Condition	Population		Contaminated		Contaminated	
	r_{it}	$r_{i\theta}$	r_{it}		$r_{i\theta}$	
			20%	30%	20%	30%
N=500, K=20	.012	.005	.012	.011	.006	.008
N=500, K=40	.007	.0025	.005	.006	.0025	.003
N=1000, K=20	.009	.006	.008	.009	.004	.004
N=1000, K=40	.0075	.004	.006	.004	.0035	.002

Note. N : the number of test takers. K: the number of items.

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

10.21831/pep.v27i2.21477

The results of the item selection process using generated data demonstrate the proportion of items eliminated using two techniques, r_{it} dan $r_{i\theta}$. A comparison of the performance of the two techniques can be inferred from the false alarm rate across various conditions, as depicted in Table 2 and Table 3. The false alarm rate, which constitutes a type 1 error, signifies instances where high-quality items are wrongly classified as low-quality and subsequently eliminated.

Table 2 illustrates that both techniques exhibit comparable performance under various conditions, as evidenced by a false alarm rate less than 5%. A closer examination of the data contamination percentage reveals that the false alarm rate of the $r_{i\theta}$ technique consistently outperforms that of the r_{it} technique. Additionally, the results indicate a negative correlation between the number of items and the false alarm rate, with an increase in the number of items leading to a decrease in the false alarm rate.

Table 3. Proportion of Eliminated Items Based on r_{it} and $r_{i\theta}$ Using Cut-off Point .30

Condition	Population		Contaminated		Contaminated	
	r_{it}	$r_{i\theta}$	r_{it}		$r_{i\theta}$	
			20%	30%	20%	30%
N=500, K=20	.045	.017	.048	.053	.059	.052
N=500, K=40	.044	.0215	.0345	.033	.035	.0225
N=1000, K=20	.043	.011	.045	.05	.041	.035
N=1000, K=40	.035	.019	.029	.028	.0345	.019

Note. N : the number of test takers. K: the number of items

Table 3 presents the findings of the evaluation of the false alarm rate for two techniques, $r_{i\theta}$ and r_{it}, under various conditions. The results indicate a negative correlation between the number of items and the false alarm rate, with an increase in the number of items leading to a decrease in the false alarm rate. On the other hand, the analysis shows that in most conditions, the false alarm rate of the $r_{i\theta}$ technique consistently outperforms that of the r_{it} technique, except under conditions of 20% contamination with N=500 and K=20. However, it is noteworthy that there are three instances in which the false alarm rate surpasses the threshold of 5%. The result indicates that when the cut-off threshold is elevated to .30, the false alarm rate experiences an increase, particularly in the scenario where N=500 and K=20. The $r_{i\theta}$ technique shows a false alarm rate that surpasses the 5% threshold when the data is contaminated by 20% and 30% of low-quality items under the same conditions. Meanwhile, the r_{it} technique demonstrates a false alarm rate of 0.053 under the conditions of N=500 and K=20, with a contamination level of 30%.

Real Data

Table 4. Proportion of Eliminated Items Based on r_{it} and $r_{i\theta}$ Using Cut-Off Points .2 and .3

Statistic	Cutoff	
	.20	.30
r_{it}	.119	.089
$r_{i\theta}$	.170	.081

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

10.21831/pep.v27i2.21477

Subsequently, a real data analysis was conducted using the item-total correlation and item-theta correlation methods for item selection. The data used in the analysis was obtained from the 2019 administration of the TIMSS Electronic Mathematics assessment in Malaysia, identified by code BSAMYSZ7. The data consisted of 135 items and a sample of 534 students from 13 schools in the country. The results of the real data analysis are presented in Table 4, which showcases the results of the item selection procedure.

According to Table 4, the results reveal that across all cut-off scenarios, the proportion of items eliminated by the $\boldsymbol{r_{i\theta}}$ technique consistently surpasses that of the $\boldsymbol{r_{it}}$ technique. The disparity between the proportions is particularly pronounced when a cut-off point of .20 is utilized in the item selection process. However, the difference in proportion becomes less pronounced when the cut-off point is increased to .30. Additionally, Figure 2 provides an illustration of the comparison between the $\boldsymbol{r_{it}}$ and $\boldsymbol{r_{i\theta}}$ techniques.

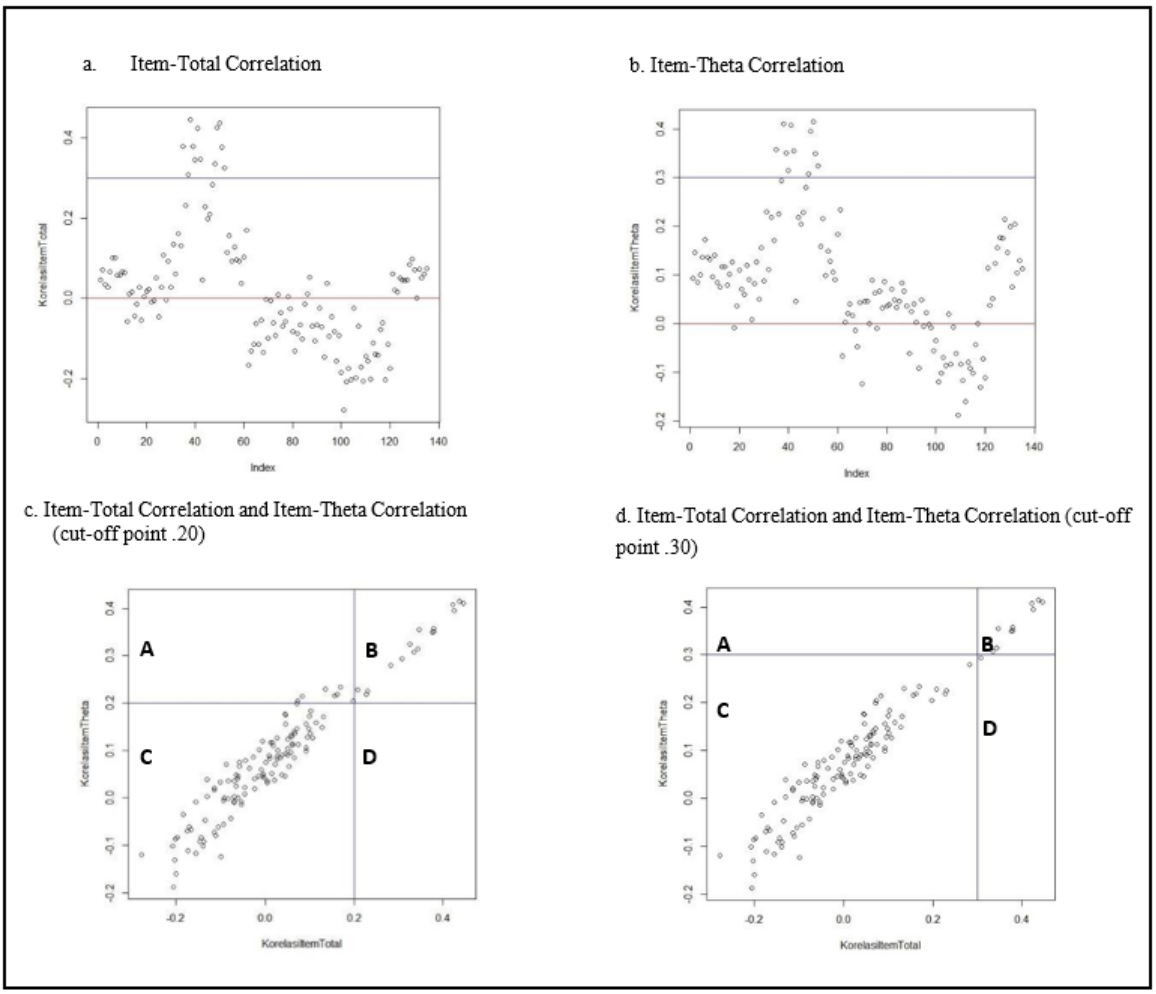

Figure 2. Distribution of Item-Total Correlation and Item-Theta Correlation Values

The results of the correlation analysis depicted in Figure 2a demonstrate that there were 73 items that displayed a positive correlation and 62 items that exhibited a negative correlation in the item-total correlation. Utilizing a cut-off point of .30, only 12 items were identified as high-quality items, specifically items 35, 37, 38, 39, 40, 41, 42, 48, 49, 50, 51, and 52. The remaining items with correlation values below .30 require further evaluation and potential improvement.

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

10.21831/pep.v27i2.21477

Figure 2b presents the results of the item-theta correlation analysis. The analysis reveals that 103 items exhibited positive correlations and 32 items exhibited negative correlations. The utilization of a cut-off point of .30 resulted in the identification of 11 high-quality items, specifically items 35, 38, 39, 40, 41, 42, 48, 49, 50, 51, and 52.

Further analysis, Figures 2c and 2d depict four designated regions: A, B, C, and D. Regions A and D are designated as "inconsistencies" reflecting the divergences in the decision-making processes between the item-total correlation and the item-theta correlation. The items located in region A are identified as high-quality items based on their item-theta correlation values but not on the item-total correlation values. Conversely, the items situated in region D are recognized as high-quality items based on their item-total correlation values, but not on the item-theta correlation values.

Subsequently, regions B and C are designated as "consistencies," indicating a concurrence between the item-total correlation and item-theta correlation in classifying items as high-quality (located in region B) or low-quality (located in region C). The items situated in region B are deemed to be of high quality due to their correlation values exceeding the specified cut-off point. Conversely, items in region C are considered for revision as their correlation values fall below the specified cut-off point.

As illustrated in Figures 2c and 2d, there are differences in the outcome based on different cutoff points. Figure 2c depicts the results when a cutoff point of .20 is utilized in the study. The results show that the correlation values of each item fall into regions A, B, and C, with 7 items in region A, 16 items in region B, and 112 items in region C. This indicates that there are 7 items in the "inconsistencies" region, which implies a discrepancy in the final decision based on the item-total correlation and the item-theta correlation when using a cutoff point of .20.

As displayed in Figure 2d, a different result was obtained when the cutoff point used was .30. The figure indicates that the regions populated with the items are regions B, C, and D, with 11 items in region B, 123 items in region C, and 1 item in region D. This indicates that when using the .30 cutoff point, 1 inconsistency in the final decision was obtained based on the correlation values of item-total and item-theta.

According to Figures 2c and 2d, it can be concluded that the utilization of the same data set can result in different final decision scenarios depending on the applied cut-off points. The findings demonstrate that discrepancies are predominantly found in items that exhibit low correlation values. This is evidenced by the higher number of items situated in area A, which represents the "inconsistencies" region, in Figure 2c. This can be attributed to the wider distribution of values in low correlation scenarios, leading to greater potential for mismatch between the two decision scenarios.

Discussion

The purpose of this study was to investigate potential differences in decision-making regarding eliminated items within a dataset using two approaches: item-total correlation and item-theta correlation. By analyzing data obtained through simulation studies and real data, the results demonstrated that there are differences and similarities in the results depending on several conditions.

Based on the results of the simulation study, it was found that the performance of $r_{i\theta}$ consistently outperformed r_{it} under all conditions, particularly when using a cut-off point of .20. Table 2 displays the false alarm rate for $r_{i\theta}$ consistently lower than that of r_{it}. When using a cut-off point of .30 (Table 3), the $r_{i\theta}$ technique exhibited better performance than r_{it} under most conditions, except for two conditions that showed an increase in the false alarm rate value.

Furthermore, several findings related to the number of items were also discovered. Specifically, there is a negative relationship between the number of items and the false alarm

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

10.21831/pep.v27i2.21477

rate under all conditions. As the number of items increased, the false alarm rate consistently decreased. These results indicate that the accuracy of both techniques in selecting the most appropriate items is significantly influenced by the number of items. Therefore, the greater the number of items included, the higher the accuracy of both techniques in selecting the best items. This aligns with the research conducted by Guo et al (2022), which asserts that measurement error decreases as the length of the test increases. A reduction in measurement error signifies high accuracy. That's why in the simulation results of this study, accuracy in item selection increases when the number of items grows.

In the simulation study, a pattern in the correlation values of both techniques, which was related to the set cut-off point was also identified. When the cut-off point was increased to .30, the false alarm rate also increased, although most of the increase remained below .05. Additionally, both techniques exhibited a decline in accuracy across all conditions when the cut-off point was raised. The item-total correlation cut-off value of 0.30 is often used to check how well an item relates to the total score. This value helps researchers see how well each item can tell the difference between different responses, and shows which items might need to be revised or eliminated. In the education field, Kesici and Tunç (2018) eliminated items that had a correlation lower than 0.30, as these had low discrimination power. This illustrates how the use of this cut-off value can effectively determine the quality of the items. Also, in nursing education, Bae et al (2023) mentioned that the appropriate range for the item-total correlation is 0.30-0.70.

The results obtained from the real data analysis indicated that when using a cut-off point of .30, the r_{it} technique selected 12 high-quality items, while the $r_{i\theta}$ technique selected 11 high-quality items. Interestingly, the findings showed a similarity in the number of items selected by both techniques, with 11 items being the exact same across both techniques. This suggests an agreement in the results obtained from both $r_{i\theta}$ and r_{it} techniques. While the correlation values for each technique exhibited differences, the ultimate decision in selecting the best items remained the same.

A significant difference in outcomes was observed when the cut-off point was lowered to .20, indicating a mismatch in selecting the best items using the $r_{i\theta}$ and r_{it} techniques. Specifically, the r_{it} technique only selected 16 high-quality items, while the $r_{i\theta}$ technique identified 23 high-quality items. The findings highlight the importance of considering the cut-off point when selecting the best items for analysis using either technique.

Based on the results, it was observed that at the cut off point of .20, a total of 7 items showed a discrepancy between r_{it} and $r_{i\theta}$. However, with a higher cut off point of .30, only 1 item demonstrated inconsistency between r_{it} and $r_{i\theta}$. These findings suggest that a strong agreement between the results of the two techniques occurs at a correlation value of .30 or higher. According to Ebel (Dichoso et al., 2020),determining the appropriate cut off point in correlation is crucial in identifying which items need to be selected, deleted, or revised. The appropriate cut off point to determine that an item can be used is .30. This assertion is consistent with Crocker and Algina (2008) that items scoring between .30 and 1 indicate high quality and should be retained. Azwar (1994) suggested that an item discrimination is considered low when its correlation value is below .30. However, it should be noted that under certain circumstances, the cut-off point of .30 may be reduced to .25.

CONCLUSION

The present study has yielded three critical findings. Firstly, it has been observed that the accuracy of the r_{it} and $r_{i\theta}$ techniques in item selection can be significantly impacted by the number of items. Secondly, the performance of both techniques has been found to be comparable for items with high discrimination power, as evidenced by a correlation value of .30

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

10.21831/pep.v27i2.21477

or higher. Finally, the study underscores the importance of determining an appropriate cut off point for correlation values in the item selection process.

ACKNOWLEDGMENT

I would like to thank you funding for this research. This work was supported by FISIP Universitas Brawijaya.

REFERENCES

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. https://doi.org/10.1007/978-981-13-7496-8

Azwar, S. (1994). seleksi item dalam penyusunan skala psikologi. *Buletin Psikologi*, *2*(2), 26–33. https://jurnal.ugm.ac.id/buletinpsikologi/article/view/13277/9500

Bae, J., Lee, J. H., Choi, M., Jang, Y., Park, C. G., & Lee, Y. J. (2023). Development of the clinical reasoning competency scale for nurses. *BMC Nursing*, *22*(1), 1–8. https://doi.org/10.1186/S12912-023-01244-6/TABLES/4

Baker, F. B. (2001). *The Basics of Item Response Theory* (2nd ed.). eric clearinghouse on assessment and evaluation.

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

Bandalos, & Deborah L. (2018). *Measurement Theory and Applications for the Social Sciences*. www.guilford.com/MSS

Bock, R. D., & Gibbons, R. D. (2021). *Item response theory* (1st ed.). John Wiley & Sons. https://doi.org/10.1002/9781119716723

Bulut, O., & Sünbül, Ö. (2017). R Programlama Dili ile Madde Tepki Kuramında Monte Carlo Simülasyon Çalışmaları. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 266–287. https://doi.org/10.21031/epod.305821

Crocker, L. M., & Algina, James. (2008). *Introduction to classical and modern test theory*. Cengage Learning. https://archive.org/details/introductiontocl00croc

de Ayala, R. J. (2022). *The theory and practice of item response theory* (second). The Guilford Press.

Desjardins, C. D., & Bulut Boca Raton, O. (2019). Handbook of Educational Measurement and Psychometrics Using R. *Https://Doi.Org/10.1080/00031305.2019.1676110*, *73*(4), 415–416. https://doi.org/10.1080/00031305.2019.1676110

Dichoso, A. A., Joy, R., & Cabauatan, M. (2020). Test Item Analyzer Using Point-Biserial Correlation And P-Values. *International Journal Of Scientific & Technology Research*, *9*(4), 2122–2126. www.ijstr.org

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

10.21831/pep.v27i2.21477

Feinberg, R. A., & Rubright, J. D. (2016). Conducting Simulation Studies in Psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/EMIP.12111

Finch, W. H., & French, B. F. (2018). Educational and Psychological Measurement. *Educational and Psychological Measurement*. https://doi.org/10.4324/9781315650951

Guo, H., Lu, R., Johnson, M. S., & McCaffrey, D. F. (2022). Alternative Methods for Item Parameter Estimation: From CTT to IRT. *ETS Research Report Series*, *2022*(1), 1–16. https://doi.org/10.1002/ets2.12355

Hu, Z., Lin, L., Wang, Y., Li, J., Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The Integration of Classical Testing Theory and Item Response Theory. *Psychology*, *12*(9), 1397–1409. https://doi.org/10.4236/PSYCH.2021.129088

Jacobs, P., & Viechtbauer, W. (2017). Estimation of the biserial correlation and its sampling variance for use in meta-analysis. *Research Synthesis Methods*, *8*(2), 161–180. https://doi.org/10.1002/JRSM.1218

Karakaya, N., & Kiliç, M. (2021). Turkish Adaptation of the Workplace Breastfeeding Support Scale: A Validity and Reliability Study. *Samsun Sağlık Bilimleri Dergisi*. https://doi.org/10.47115/jshs.1029188

Kesici, A., & Tunç, N. F. (2018). The Development of the Digital Addiction Scale for the University Students: Reliability and Validity Study. *Universal Journal of Educational Research*, *6*(1), 91–98. https://doi.org/10.13189/ujer.2018.060108

Luecht, R. M., & Hambleton, R. K. (2021). Item Response Theory : A Historical Perspective and Brief Introduction to Applications. *The History of Educational Measurement*, 232–262. https://doi.org/10.4324/9780367815318-11

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

Paek, Insu, Cole, & Ki. (2019). *USING R FOR ITEM RESPONSE THEORY MODEL APPLICATIONS*.

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *Type Package Title Test Analysis Modules*. https://doi.org/10.1177/0146621697211001

Supratiknya, A. (2014). *Pengukuran Psikologis* (1st ed.). Universitas Sanata Dharma.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational Measurement for Applied Researchers. *Educational Measurement for Applied Researchers*. https://doi.org/10.1007/978-981-10-3302-5

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

10.21831/pep.v27i2.21477

Xie, D., & Cobb, C. L. (2020). Item Analysis. *The Wiley Encyclopedia of Personality and Individual Differences*, 159–163. https://doi.org/10.1002/9781118970843.CH97

Yilmaz, M., & Keskin, H. (2020). Is a Universal Model of a "Good" National Education System that brings Economic Returns Emerging? *Anadolu Üniversitesi Sosyal Bilimler Dergisi*, *20*(Özel Sayı), 61–72. https://doi.org/10.18037/ausbd.725563

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