Jurnal Penelitian dan Evaluasi Pendidikan
Volume 28, No. 2, December 2024 (161-172)
Online: http://journal.uny.ac.id/index.php/jpep

# Psychometric quality of multiple-choice tests under classical test theory (CTT): AnBuso, Iteman, and R

**Siti Nurjanah[1*]; Muhammad Iqbal[1]; Zafrullah Zafrullah[1]; Muhammad Naim Mahmud[2]; D'aquinaldo Stefanus Fani Seran[1]; Izzul Kiram Suardi[1]; Lovieanta Arriza[1]**
[1]Universitas Negeri Yogyakarta, Indonesia
[2]Universitas Muslim Buton, Indonesia
[*]Corresponding Author. E-mail: siti960pasca.2023@student.uny.ac.id

**ARTICLE INFO**

**Article History**
**Submitted:**
23 February 2024
**Revised:**
04 July 2024
**Accepted:**
20 August 2024

**Keywords**
anbuso; iteman; rstudio; classical test theory; psychometric quality analysis

**Scan Me:**

**ABSTRACT**

Psychometric quality analysis of psychological instruments was important to ensure credible measurement. This study aims to compare the psychometric quality analysis of multiple-choice test items using three different applications to evaluate the advantages and disadvantages of the features provided in supporting classical test theory analysis. This study used a quantitative approach by analysing dichotomous data from 50 participants of a 30-item multiple-choice test. The data were analysed using three applications (AnBuso, Iteman, and R) to compare the statistical output of the main psychometric parameters of the classical test theory, such as difficulty index, discrimination index, and distractor effectiveness. Data analysis was conducted descriptively and quantitatively by comparing the features provided by the application in support of classical test theory analysis to evaluate the advantages and disadvantages of each application. The study found that all three applications produced similar results for the difficulty index, distractor effectiveness, and discrimination index. AnBuso proved user-friendly but limited in capacity, Iteman offered comprehensive output with restricted free functionality, and R provided flexibility but required programming expertise. The application demonstrated unique strengths that are suitable for different research needs and user proficiencies. The choice of application should consider factors such as analysis complexity, sample size, and user expertise. Further research into paid options and diverse test conditions is recommended for a more comprehensive evaluation of these applications in classical test theory analysis.



**To cite this article (in APA style):**
Nurjanah, S., Iqbal, M., Zafrullah, Z., Mahmud, M. N., Seran, D. S. F., Suardi, I. K., & Arriza, L. (2024). Psychometric quality of multiple-choice tests under classical test theory (CTT): AnBuso, Iteman, and R. *Jurnal Penelitian dan Evaluasi Pendidikan, 28(2)*, 161-172 doi: https://doi.org/10.21831/pep.v28i2.71542

## INTRODUCTION

Psychometric quality needs to be analysed. The measurement theory framework used had a significant impact. One framework often used in psychometric analysis is Classical Test Theory (CTT) (Awopeju & Afolabi, 2016; Ayanwale et al., 2022; Siegert et al., 2022). CTT provides a conceptual foundation that views test scores as the result of two main factors: individual ability and item difficulty (Amiruddin & Langamin, 2022; Mariana et al., 2023; Subali et al., 2021). CTT provides an intuitive and easy-to-understand framework. the Model has advantages, especially in taking into account aspects such as discrimination and distractor effectiveness (Abedalaziz & Leng, 2013; Ashraf & Author, 2020; Shanmugam & Rajoo, 2020). Therefore, for the quality of the instrument, it was important to understand the additional parameters that include the discrimination index and the effectiveness of the distractors.

The difficulty index was measured through the proportion of participants who could answer the item correctly, giving an idea of how difficult or easy an item was for participants (Aybek, 2023; Cappelleri et al., 2014; Setiawati et al., 2023). Discrimination index, as a key


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

10.21831/pep.v28i2.71542

parameter in CTT, reflects the ability of a test item to distinguish between participants with different ability levels; a high discrimination index indicates the effectiveness of the item in distinguishing individuals with diverse ability levels (Awopeju & Afolabi, 2016; Eleje et al., 2019; Hu et al., 2021). Meanwhile, the effectiveness of distractors, which measures the extent to which incorrect answer options can distinguish between competent and incompetent individuals, was key in accurately evaluating participants' responses (Chauhan et al., 2023; Pratama, 2019; Rodriguez et al., 2014). Understanding and optimising these parameters in CTT-based psychometric quality analysis could improve the validity of the instrument, ensuring that it provides accurate and useful information about the psychological characteristics of the individual being measured.

One of the local applications in Indonesia that was quite interested in item calibration applications using classical test theory was AnBuso. AnBuso was an item analysis program that was simplified to assist teachers in preparing administrative reports related to item analysis using Excel (Muhson, 2017; Yuwono et al., 2020). AnBuso was a classic test theory application that used features, functions, and formulas available in Microsoft Excel, so it could only operate within the platform (Muhson et al., 2017). So, in practicality, the application was used in item analysis, especially in classical test theory. There were also other applications that could be used as a tool for item analysis, namely Iteman.

Another application that was widely used for instrument analysis with classical test theory was Iteman. Iteman was an analysis program that has been part of the assessment systems corporation's item and test analysis package since 1981, which focuses on generating item and participant statistics based on Classical Test Theory (CTT) (Jatnika et al., 2020; Monamodi, 2016). Iteman was an application of classical test theory adopted to calibrate item indices such as difficulty index, discrimination index, and distractor strength on each exam item (Kirya et al., 2023; Shakir et al., 2022). The application has quite a few complete features. There was another application that could be used as an option in conducting CTT analysis, namely the R Program.

The R program was an application specifically designed for item analysis using the R language, which could provide various features and functionality that support the item analysis process efficiently and effectively (Arriza et al, 2024; Shahmirzadi, 2023; Travezaño-Cabrera et al., 2022). The R Studio software interface from R Program was an integrated development environment (IDE) popular in item analysis and statistical programming. The R Studio can write, edit, and run R scripts to perform data analysis, such as analyzing tests, test items, data manipulation, statistical modelling, and visualization of results. The application has a user-friendly interface and provides various additional tools and packages useful in the item analysis process. The application has a user-friendly interface and provides various additional tools and packages useful in the item analysis process (Am et al., 2023; Marfu'ah et al., 2023). The R Studio is a software interface that uses the R programming language. It has features and functionality that support the data analysis process efficiently and effectively, as well as providing a friendly user interface and additional useful instruments. The three applications for analyzing the quality of instruments with CTT were necessary to make a comparison between them.

Previous research by Pradani and Efendi (2023) explored the quality of junior high school examination items in the Rembang District, focusing on critical elements such as content validity, reliability, difficulty index, discrimination index, and distractors. The findings indicated variations in difficulty index and mixed-item quality. In addition, Irawati et al. (2020) conducted a study using AnBuso to evaluate question items with the participation of 200 students. The analysis includes aspects of the discrimination index, difficulty index, and distractor effectiveness. In the results, 45.71% of the items were rated as having good quality, providing support for the usefulness of AnBuso in analysing classical test theory. Although studies have


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

10.21831/pep.v28i2.71542

been conducted to analyse the item quality of instruments under the CTT approach using specific applications, a study simultaneously comparing three applications has yet to be conducted.

This study will compare three applications for analysing items under the classical test theory approach. Specifically, this study would compare the results of assessing the psychometric quality of a test under the CTT approach from three applications based on the parameters of discrimination index, difficulty index, and distractor effectiveness. We would also analyse the advantages and disadvantages of the three applications in supporting the analysis of the psychometric quality of an item under the CTT approach by evaluating their features so as to provide a comprehensive picture of their reliability and effectiveness in the context of psychometric assessment.

## RESEARCH METHOD

This research used a quantitative approach to achieve its goals by analysing dichotomous data. The data used was obtained from secondary sources, namely the results of research conducted by Mahmud (2021). The researcher has obtained permission from the data owner to use it in this study. The data came from a multiple-choice test instrument consisting of 30 items with four alternative answers and involving 50 test participants. The number of test participants of 50 was chosen because it meets the maximum number that could be analysed by AnBuso, Iteman (in free license version), and R (Guyer & Thompson, 2013; Muhson et al, 2013). The raw data was then adjusted to the input needs of each application. There were three types of input data prepared because the data would be analysed by three applications that have different characteristics of input data types. Then, the test items were evaluated for psychometric quality using the classical test theory approach with the help of AnBuso version 8, Iteman version 4.3, and R version 4.3.1.

Table 1. Interpretation of Item difficulty and Discrimination Indeces

| Index | Value | Interpretation |
|---|---|---|
| Difficulty Index ($p$) | $p \leq 0.30$ | Difficult |
| | $0.31 \leq p \leq 0.70$ | Moderately difficult |
| | $p > 0.70$ | Easy |
| Discrimination Index ($D$) | $D \geq 0.40$ | Item is functioning quite satisfactorily |
| | $0.30 \leq D \leq 0.39$ | Good item; little or no revision is required |
| | $0.20 \leq D \leq 0.29$ | Item is marginal and need revision |
| | $D \leq 0.19$ | Poor item; should be eliminated or completely revised |

This study focuses on comparing the statistical output generated by three different applications with the specific version. The researcher would compare common parameters of item psychometric quality from the classical test theory approach, such as difficulty index, discrimination index, and distractor effectiveness (Shakurnia et al., 2022). The justification of an item's difficulty index was based on the standards proposed by Henning (1987), while the discrimination index of an item uses the standards proposed by Ebel and Frisbie (1991). These qualifications are outlined in Table 1. Furthermore, to be classified as functioning, a distractor must be selected by at least 5% of test takers (Gierl et al., 2017; Raymond et al., 2019; Rogausch et al., 2010; Sajjad et al., 2020; Tarrant et al., 2009). With a total of 50 test takers, an answer alternative was said to be functional if it was selected by at least three test takers. Data analysis would be done descriptively and quantitatively.


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

10.21831/pep.v28i2.71542

Then, the researcher would compare the features provided by each application to support the analysis of the psychometric quality of an item using the CTT approach. From this analysis, the advantages and disadvantages of each application would be evaluated. Figure 1 shows the series of research processes, starting from the data search stage to the final analysis carried out in this study.

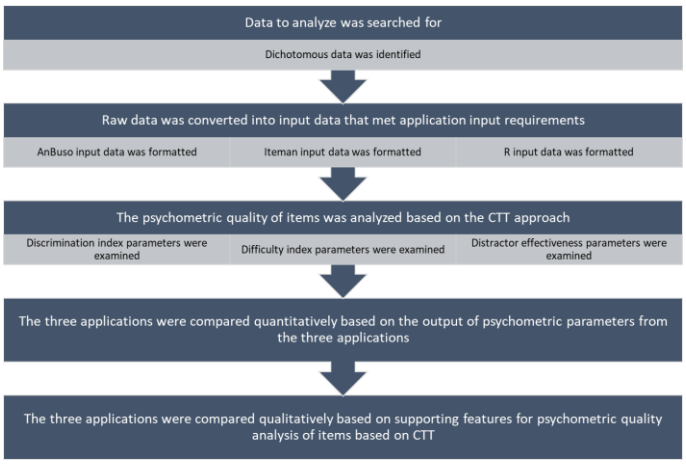


Figure 1. Research process flow to compare AnBuso, Iteman, and R for item psychometric quality analysis based on CTT

## FINDINGS AND DISCUSSION

Analysing the psychometric quality of tests using the CTT approach (parameters of discrimination index, difficulty index, and effectiveness of distractors). Table 2 presents detailed information on the difficulty index for each item based on the three applications. In this study was not significant difference in the difficulty index parameter between the three applications used. This result was consistent with the Classical Test Theory (CTT) approach. Difficulty was measured through the ratio of correct answers to the maximum score of a test (Shakurnia et al., 2022).

Table 2. Difficulty index according to AnBuso, Iteman, and R applications.

| Number of item | Difficulty Index | | | Number of item | Difficulty Index | | |
|---|---|---|---|---|---|---|---|
| | AnBuso | Iteman | R | | AnBuso | Iteman | R |
| 1 | 0.88 | 0.88 | 0.88 | 16 | 0.94 | 0.94 | 0.94 |
| 2 | 0.96 | 0.96 | 0.96 | 17 | 0.66 | 0.66 | 0.66 |
| 3 | 0.94 | 0.94 | 0.94 | 18 | 0.72 | 0.72 | 0.72 |
| 4 | 0.98 | 0.98 | 0.98 | 19 | 0.68 | 0.68 | 0.68 |
| 5 | 0.90 | 0.90 | 0.90 | 20 | 0.76 | 0.76 | 0.76 |
| 6 | 0.76 | 0.76 | 0.76 | 21 | 0.96 | 0.96 | 0.96 |
| 7 | 0.90 | 0.90 | 0.90 | 22 | 0.94 | 0.94 | 0.94 |
| 8 | 0.80 | 0.80 | 0.80 | 23 | 0.94 | 0.94 | 0.94 |
| 9 | 0.68 | 0.68 | 0.68 | 24 | 0.76 | 0.76 | 0.76 |
| 10 | 0.56 | 0.56 | 0.56 | 25 | 0.36 | 0.36 | 0.36 |
| 11 | 0.86 | 0.86 | 0.86 | 26 | 0.34 | 0.34 | 0.34 |
| 12 | 0.98 | 0.98 | 0.98 | 27 | 0.76 | 0.76 | 0.76 |
| 13 | 0.82 | 0.82 | 0.82 | 28 | 0.62 | 0.62 | 0.62 |
| 14 | 0.68 | 0.68 | 0.68 | 29 | 0.54 | 0.54 | 0.54 |
| 15 | 0.76 | 0.76 | 0.76 | 30 | 0.26 | 0.26 | 0.26 |


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

10.21831/pep.v28i2.71542

The three apps calculated the number of correct answers as well as the maximum score of a test exactly the same, given that the process involves simple calculations. Thus, the output of the difficulty parameter did not any significant difference among the three apps. This finding indicated that for evaluating the difficulty index parameter, test developers could use these three apps without any hesitation. Referring to the difficulty index proposed by Henning (1987), the analysis of the investigated tests showed that one item was categorised as difficult, nine items as moderately difficult, and twenty items as easy. Items classified as easy could be eliminated or placed at the beginning of the test, which serves as a warm-up. Meanwhile, items classified as difficult require re-analysis related to language, controversy, or errors that may occur in the item. Decision-making regarding the inclusion of items classified as difficult in the test depends on the purpose of the measurement being carried out (Hingorjo & Jaleel, 2012).

Table 3 provides a detailed overview of the ineffective answer alternatives for each item across the three applications.

Table 3. Ineffective answer alternatives according to AnBuso, Iteman, and R applications.

| Number of item | Ineffective Distractors | | | Number of item | Ineffective Distractors | | |
|---|---|---|---|---|---|---|---|
| | AnBuso | Iteman | R | | AnBuso | Iteman | R |
| 1 | - | AD | AD | 16 | C | BCD | BCD |
| 2 | A | ACD | ACD | 17 | - | B | B |
| 3 | AD | AD | AD | 18 | - | - | - |
| 4 | BD | ABD | ABD | 19 | - | A | A |
| 5 | B | AB | AB | 20 | - | B | B |
| 6 | - | B | B | 21 | D | ABD | ABD |
| 7 | - | BCD | BCD | 22 | - | BCD | BCD |
| 8 | - | A | A | 23 | AC | AC | AC |
| 9 | - | D | D | 24 | - | D | D |
| 10 | - | - | - | 25 | - | - | - |
| 11 | - | B | B | 26 | - | - | - |
| 12 | AD | ACD | ACD | 27 | - | B | B |
| 13 | - | A | A | 28 | A | A | A |
| 14 | - | A | A | 29 | - | - | - |
| 15 | - | A | A | 30 | - | - | - |

There was a significant similarity in the Iteman and R output results on the parameter of distractor effectiveness, which can be caused by the decision-making that was still done manually by the researcher. The researcher employed a criterion of a minimum 5% selection rate for an answer alternative to be considered effective. Both applications showed similarities in the output because the analysis of the number of correct answers, the number of test takers who chose an alternative answer, and the number of wrong answers were the same in each application. In addition, there was a slight difference in the output produced by AnBuso. This was due to the use of different qualifications in determining the effectiveness of an alternative answer.

Further examination of AnBuso's results revealed that it employs a less stringent criterion, considering an alternative answer ineffective only when no test participants select it (indicated by a 0% distribution). While the percentage figures were identical across all three applications, the variation in evaluation judgments stemmed from these differing criteria. It was crucial to note that the choice of criteria ultimately depends on the specific requirements of researchers or practitioners. AnBuso's automatic judgment feature proves particularly advantageous for validating teacher-made tests in educational settings, as it allows for a more flexible approach to test validation. These benefits align with the application's intended purpose and target users (Muhson et al, 2013).


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

10.21831/pep.v28i2.71542

In contrast, Iteman and R require additional configuration to produce judgments, unlike AnBuso's automated output. Researchers using these applications must establish their own criteria for ineffective distractors, necessitating a more in-depth theoretical understanding. While R could be configured to generate automatic judgments similar to AnBuso, this requires syntax modifications to align with specific research criteria. This may pose challenges for users who are less proficient in syntax manipulation. However, this limitation can also be viewed as an advantage, offering researchers greater flexibility in making judgments tailored to their studies.

Table 4 presents a comprehensive breakdown of discrimination indices for each item in the third application. It should be noted the discrimination index analyzed in this study was a point-biserial correlation (pubis). Pbis is generally preferred to biserial correlation (bis) for item discrimination in classical test theory because of its greater robustness to violations of normality assumptions and a more accurate representation of the relationship between item performance and overall test scores, particularly for dichotomous items. In addition, pbis is easier to calculate and interpret, does not require complex distribution assumptions, and can be widely applied in various testing contexts, providing researchers with a reliable and practical tool for measuring the quality of test questions (Das & Richman, 2022; Shen et al., 2023). Thus, the use of pbis can increase accuracy and efficiency in evaluating the quality of test items in various testing settings.

Table 4. Discrimination index, difficulty index, and ineffective answer alternatives according to AnBuso, Iteman, and R applications.

| Number of item | Discrimination Index | | | classification | Number of item | Discrimination Index | | | classification |
|---|---|---|---|---|---|---|---|---|---|
| | AnBuso | Iteman | R | | | AnBuso | Iteman | R | |
| | pbis | Rpbis | pbis | | | pbis | Rpbis | pbis | |
| 1 | 0.35 | 0.35 | 0.35 | Good item | 16 | 0.20 | 0.20 | 0.20 | Marginal |
| 2 | -0.20 | -0.20 | -0.20 | Poor item | 17 | 0.34 | 0.34 | 0.34 | Good item |
| 3 | 0.20 | 0.20 | 0.20 | Marginal | 18 | 0.06 | 0.06 | 0.06 | Poor item |
| 4 | 0.09 | 0.09 | 0.09 | Poor item | 19 | 0.07 | 0.07 | 0.07 | Poor item |
| 5 | 0.00 | 0.00 | 0.00 | Poor item | 20 | 0.06 | 0.06 | 0.06 | Poor item |
| 6 | 0.33 | 0.33 | 0.33 | Good item | 21 | 0.02 | 0.02 | 0.02 | Poor item |
| 7 | 0.25 | 0.25 | 0.25 | Marginal | 22 | 0.01 | 0.01 | 0.01 | Poor item |
| 8 | -0.11 | -0.11 | -0.11 | Poor item | 23 | -0.05 | -0.05 | -0.05 | Poor item |
| 9 | 0.29 | 0.29 | 0.29 | Marginal | 24 | 0.02 | 0.02 | 0.02 | Poor item |
| 10 | 0.23 | 0.24 | 0.23 | Marginal | 25 | 0.32 | 0.33 | 0.32 | Good item |
| 11 | 0.24 | 0.24 | 0.24 | Marginal | 26 | 0.14 | 0.14 | 0.14 | Poor item |
| 12 | -0.01 | -0.02 | -0.01 | Poor item | 27 | -0.15 | -0.15 | -0.15 | Poor item |
| 13 | -0.02 | -0.02 | -0.02 | Poor item | 28 | 0.28 | 0.28 | 0.28 | Marginal |
| 14 | 0.08 | 0.09 | 0.08 | Poor item | 29 | -0.11 | -0.11 | -0.11 | Poor item |
| 15 | -0.23 | -0.23 | -0.23 | Poor item | 30 | 0.34 | 0.34 | 0.34 | Good item |

The analysis showed consistent discrimination index (pbis) values across the three applications. Marginal differences were detected in some items, such as items 10 and 12, with deviations of around 0.01. However, the differences were considered negligible and didn't have a significant impact on the interpretation of item quality. Therefore, for researchers who want to analyze discriminability parameters within the framework of Classical Test Theory (CTT) using the point-biserial method, the choice between Anbuso, Iteman, and R can be considered equivalent. The discrimination index (pbis) measures the extent to which an item can differentiate between participants with high and low ability (Cobbinah & Ntumi, 2022; Setiawati et al., 2023). A high pbis value indicates that the item was effective in discrimination, while a low value indicates the opposite. The absence of substantial differences in output indicates that


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

10.21831/pep.v28i2.71542

the three applications most likely apply similar algorithms or formulas for calculating pbis values. The consistency of analysis results across these three applications gives researchers flexibility in choosing the analysis tool that best suits their needs and preferences.

***The advantages and disadvantages of the application in terms of features to help analysed the psychometric quality of an item under the CTT approach***

This comparative study would be incomplete if it only considered the output of the three common parameters in Classical Test Theory (CTT). The selection of an application to analyse item quality was not solely based on these considerations. In addition, various factors were the main considerations in determining which analysis application to use. The factors such as available features, accessibility, cost, and others were also an essential part of the decision-making process. Therefore, the research would conduct a comprehensive review of these three applications by considering the support provided by each application in analysing the psychometric quality of an item according to the classical test theory approach. Table 5 provides information on the strengths and weaknesses of each application.

Table 5. Advantages and disadvantages of AnBuso, Iteman, and R for item psychometric quality analysis with the CTT approach.

| | AnBuso | Iteman | R |
|---|---|---|---|
| Advantages | • Open-source<br>• Practically used for beginners.<br>• Basic outputs were available that were suitable for practical classroom evaluation purposes. | • Output was very neat in a word file<br>• Quantile plots of items available.<br>• There were many statistical parameter outputs for item quality tests. | • Open-source<br>• Not friendly for beginners.<br>• Available Item Characteristic Curve (ICC) plots. |
| Disadvantages | • The number of items was limited to 50 multiple-choice questions and 10 essay questions. The number of examinees was limited to 200 people.<br>• Output statistical parameters for the item quality test were not as complex as those of Iteman and R. | • The free or basic version of Iteman software has restricted functionality, and users need to purchase a paid license to access the full set of features<br>• Need high rigor when writing input. | • We need high rigor to build syntactic.<br>• The file output of R was not as good as the output generated by the item. |

Table 5 presents the advantages and disadvantages of each application used for psychometric quality analysis. These findings were in line with those of previous studies (Berk & Griesemer, 1976; Mclaughlin, 2015; Sheng, 2019). However, it was important to realise that each application has its own characteristics beyond what has been discussed in this study. The main focus of this article was on the strengths and weaknesses of apps related to supporting psychometric analyses using classical test theory.

The presence of advantages and disadvantages in each application was a necessity. The decision to choose a particular application for test validation by test developers needs to be adjusted to the needs of psychometric quality analysis as well as the ability of individual test developers to master the available applications. The psychometric quality analysis using the classical test theory approach in the context of research has been the most recommended choice. The reasons behind this recommendation were as follows: Firstly, the result parameters of


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

10.21831/pep.v28i2.71542

Iteman were not different from those of R or AnBuso. Secondly, the output files were complex and neatly structured in Word format. Thirdly, although the statistics generated were quite complex, the use of Iteman was still relatively easy to understand. However, if the test developer has a number of items and test takers that exceed 50 and wants an open-source application, then using R is a more appropriate choice. In addition, AnBuso was a suitable option for those who were not familiar with using computer programs and did not require complex statistical output.

This study focuses specifically on applications with free license modes and predetermined test parameters. While this approach provides valuable insights for researchers and educators working within these constraints, it also presents a limitation. Future studies could expand upon this research by examining paid software options and exploring a wider range of test conditions to enhance the generalizability of these findings across diverse contexts and licensing models.

## CONCLUSION

The investigation revealed three applications produced identical difficulty index values, demonstrating consistency in this parameter's calculation. Regarding distractor effectiveness, Iteman and R showed highly similar results. At the same time, AnBuso exhibited slight variations due to its distinct criteria for determining distractor efficiency in terms of discrimination index, AnBuso, Iteman, and R generated identical point-biserial correlation (pbis) values. Each application demonstrated unique strengths and limitations. AnBuso proved user-friendly for beginners and was open-source, but it was restricted in the number of items and test participants it could analyze. Iteman produced comprehensive and well-organized output, though its free version has limited functionality. R offered high flexibility and open-source, but it requires programming expertise and generates less polished output compared to Iteman. These characteristics influence their suitability for different contexts and user proficiencies. Based on these findings, Iteman was recommended for research purposes due to its comprehensive and easily interpretable output. R was suitable for analyses involving a large number of items and test participants and for users preferring an open-source solution. AnBuso was ideal for novice users or those not requiring complex statistical output. While the three applications demonstrated comparable performance in psychometric analysis using the CTT approach, their primary differences lie in additional features, accessibility, and suitability for various levels of user expertise. The choice of application should be tailored to the specific needs of researchers or test developers, taking into account factors such as the required complexity of analysis, the number of items and test participants, and the user's proficiency level with the application. This study provides valuable insights for informed decision-making in psychometric tool selection while acknowledging the need for further research into paid software options and a broader range of test conditions to enhance the generalizability of these findings.

## REFERENCES

Abedalaziz, N., & Leng, C. H. (2013). The relationship between CTT and IRT approaches in analyzing item characteristics. *Malaysian Online Journal of Educational Sciences*, *1*(1), 64–70.

Am, M. A., Setiawati, F. A., Hadi, S., & Istiyono, E. (2023). Psychometric properties career of commitment instrument using classical test theory and graded response model. *Journal of Pedagogical Sociology and Psychology*, *5*(2), 26–40.

Amiruddin, B. J., & Langamin, M. A. (2022). Ability estimation using the classical test theory and three-parameter item response theory model. *Psych Educ*, *September*. https://doi.org/10.5281/zenodo.7063805


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

10.21831/pep.v28i2.71542

Arriza, L., Retnawati, H., & Ayuni, R. T. (2024). Item analysis of high school specialization mathematics exam questions with item response theory approach. *Barekeng: Journal of Mathematics and Its Application*, *18*(1), 151–162. https://doi.org/10.30598/barekengvol18iss1pp0151-0162

Ashraf, Z. A., & Author, C. (2020). Classical and modern methods in item analysis of test tools. *International Journal of Research and Review (Ijrrjournal.Com)*, *7*(5), 5.

Awopeju, O., & Afolabi, E. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal*, *12*. https://doi.org/10.19044/esj.2016.v12n28p263

Ayanwale, M. A., Chere-Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory: The status of the framework at the examination council of lesotho. *International Journal of Learning, Teaching and Educational Research*, *21*(8), 384–406. https://doi.org/10.26803/ijlter.21.8.22

Aybek, E. C. (2023). The relation of item difficulty between classical test theory and item response theory: Computerized adaptive test perspective. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(2), 118–127. https://doi.org/10.21031/epod.1209284

Berk, R. A., & Griesemer, H. A. (1976). Iteman: An item analysis program for tests, questionnaires, and scales. *Educational and Psychological Measurement*, *36*(1), 189–191. https://doi.org/10.1177/001316447603600122

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Chauhan, G. R., Chauhan, B. R., Vaza, J. V, & Chauhan, P. R. (2023). Relations of the number of functioning distractors with the item difficulty index and the item discrimination power in the multiple choice questions. *Cureus*, *15*(7). https://doi.org/10.7759/cureus.42492

Cobbinah, A., & Ntumi, S. (2022). Difficulty, Discrimination and Pseudo-Guessing Indices of West African Examinations Council Core Mathematics Multiple Choice Items: Theoretical and Practical Implications of Using Item Response Theory. *Journal of Research in Educational Sciences*, *13*(15), 51–60.

Das, R. R., & Richman, R. (2022). The development and application of a public energy literacy instrument. *Canadian Journal of Science, Mathematics and Technology Education*, *22*(1), 42–67.

Ebel, R. L., & Frisbie, D. A. (1991). *Frisbie, essentials of educational measurement 5th edition*. Prentice-Hall.

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2019). Comparative study of classical test theory and item response theory using item analysis results of quantitative chemistry achievement test. *The African Journal of Behavioural and Scale Development Research*, *1*(1), 26–36. https://doi.org/10.58579/ajb-sdr/1.1.2019.26

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. In *Review of Educational Research* (Vol. 87, Issue 6). https://doi.org/10.3102/0034654317726529


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

10.21831/pep.v28i2.71542

Guyer, R., & Thompson, N. A. (2013). *User's manual for Iteman 4.3* (Issue June). Assessment Systems Corporation.

Henning, G. (1987). *A guide to languange testing: Development evaluation research*. CreateSpace Independent Publishing Platform.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Hu, Z., Lin, L., Wang, Y., & Li, J. (2021). The integration of classical testing theory and item response theory. *Psychology*, *12*, 1397–1409. https://doi.org/10.4236/psych.2021.129088

Irawati, R., Ekawati, E. Y., & Budiawanti, S. (2020). Analisis butir soal ujian akhir semester gasal menggunakan program Anbuso di SMA Negeri 1 Boyolali tahun ajaran 2019/2020 (Analysis of odd semester final exam questions using the Anbuso program at SMA Negeri 1 Boyolali 2019/2020 academic year). *Jurnal Materi Dan Pembelajaran Fisika*, *10*(1), 11. https://doi.org/10.20961/jmpf.v10i1.42084

Jatnika, R., Purwono, U., Djunaidi, A., & Haffas, M. (2020). The effect of psychometric analysis series training on the item analysis ability of high school teachers in Bandung. *Journal of Physics: Conference Series*, *1477*(4), 42052.

Kirya, K. R., Mashood, K. K., & Yadav, L. L. (2023). Development of a circular motion concept inventory for use in ugandan science education. *Journal of Turkish Science Education*, *20*(1).

Mahmud, M. N. (2021). *Diagnostik kesulitan belajar Matematika siswa SMP kelas VIII di Kota Baubau menggunakan soal-soal model TIMSS (Diagnostics of mathematics learning difficulties for class VIII junior high school students in Baubau City using TIMSS model questions)*. Yogyakarta State University.

Marfu'ah, S., Masrukan, M. S., & Walid, S. P. (2023). *Analysis of mathematical reasoning ability in view of self confidence in the project based learning model with performance assessment*.

Mariana, M., Lessy, D., Riaddin, D., Hardiansyah, M. R., & Pary, C. (2023). Analysis of item characteristics of natural sciences national examinations for junior high school based on the classical test theory approach. *Jurnal Penelitian Pendidikan IPA*, *9*(10), 7837–7844. https://doi.org/10.29303/jppipa.v9i10.3698

Mclaughlin, D. (2015). ITEMAN : An item analysis and scoring program. *Applied Psychological Measurement*, *6*(1), 2015. https://doi.org/10.1177/014662168200600105

Monamodi, K. E. E. (2016). The invariance of Item Response Theory (IRT) parameter estimates and Classical Test Theory (CTT) statistics. *International Journal of Research in Social Sciences*, *6*(8), 715–737.

Muhson, Ali,. Lestari, Barkah,. Supriyanto., &, Baroroh, K. (2013). *Pengembangan software AnBuso sebagai solusi alternatif bagi guru dalam melakukan analisis butir soal secara praktis dan aplikatif*.

Muhson, A. (2017). *Penggunaan AnBuso (analisis butir soal) versi 8.0 (Use of AnBuso (question item analysis) version 8.0)*. Yogyakarta: Universitas Negeri Yogyakarta.

Muhson, A., Lestari, B., & Baroroh, K. (2017). The development of practical item analysis program for Indonesian teachers. *International Journal of Instruction*, *10*(2), 199–210.


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

10.21831/pep.v28i2.71542

Pradani, R. A., & Efendi, A. (2023). Analysis of school exam questions using the Iteman program. *Indonesian Language Education and Literature*, *8*(2), 275. https://doi.org/10.24235/ileal.v8i2.11002

Pratama, D. (2019). Analysis of Clasical Test Theory (CTT) approach on academic ability test instrument. *Jisae: Journal of Indonesian Student Assesment and Evaluation*, *5*(2), 43–54. https://doi.org/10.21009/jisae.052.05

Raymond, M. R., Stevens, C., & Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education*, *24*(1), 141–150. https://doi.org/10.1007/s10459-018-9855-9

Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor functioning in modified items for test accessibility. *SAGE Open*, *4*(4). https://doi.org/10.1177/2158244014553586

Rogausch, A., Hofer, R., & Krebs, R. (2010). Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: A simulation and survey. *BMC Medical Education*, *10*(1). https://doi.org/10.1186/1472-6920-10-85

Sajjad, M., Iltaf, S., & Khan, R. A. (2020). Nonfunctional distractor analysis: An indicator for quality of multiple choice questions. *Pakistan Journal of Medical Sciences*, *36*(5), 982–986. https://doi.org/10.12669/pjms.36.5.2439

Setiawati, F. A., Amelia, R. N., Sumintono, B., & Purwanta, E. (2023). Study item parameters of classical and modern theory of differential aptitude test: is it comparable? *European Journal of Educational Research*, *12*(2).

Shahmirzadi, N. (2023). Validation of a language center placement test: Differential item functioning. *International Journal of Language Testing*, *13*(1), 1–17.

Shakir, M. A., Shafiq, F., & Khalid, M. N. (2022). Assessment of learning achievement of visually impaired children at primary level. *Pakistan Journal of Educational Research and Evaluation (PJERE)*, *9*(2).

Shakurnia, A., Ghafourian, M., Khodadadi, A., Ghadiri, A., Amari, A., & Shariffat, M. (2022). Evaluating functional and non-functional distractors and their relationship with difficulty and discrimination indices in four-option multiple-choice questions. *Education in Medicine Journal*, *14*(4), 55–62. https://doi.org/10.21315/eimj2022.14.4.5

Shanmugam, S. K. S., & Rajoo, M. (2020). Examining the quality of english test items. *Malaysian Journal of Learning and Instruction*, *17*(2), 63–101.

Shen, Y., Lei, P.-W., & Crosson, A. C. (2023). Measuring derivational awareness for Chinese-speaking adolescents. *Research Methods in Applied Linguistics*, *2*(1), 100039.

Sheng, Y. (2019). CTT Package in R. *Measurement*, *17*(4), 211–219. https://doi.org/10.1080/15366367.2019.1600839

Siegert, R. J., Krägeloh, C. U., & Medvedev, O. N. (2022). Classical test theory and the measurement of mindfulness. In *Handbook of Assessment in Mindfulness Research* (pp. 1–14). Springer International Publishing. https://doi.org/10.1007/978-3-030-77644-2_3-1

Subali, B., Yogyakarta, U. N., Surakarta, U. M., Aminah, N. S., & Maret, U. S. (2021). *Modern test theory*. *14*(1), 647–660.


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

10.21831/pep.v28i2.71542

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, *9*(1), 1–8. https://doi.org/10.1186/1472-6920-9-40

Travezaño-Cabrera, A., Vilca, L. W., Quiroz-Becerra, J., Huerta, S. L., Delgado-Vallejos, R., & Caycho-Rodríguez, T. (2022). Meaning of Life Questionnaire (MLQ) in peruvian undergraduate students: Study of its psychometric properties from the perspective of classical test theory (CTT). *BMC Psychology*, *10*(1), 206.

Yuwono, M. R., Aribowo, E. K., Firmansah, F., & Indrayanto, B. (2020). Pelatihan AnBuso, zipgrade, dan google form sebagai alternatif penilaian pembelajaran di era digital (AnBuso, Zipgrade and Google Form training as alternative learning assessments in the digital era). *Jurnal Pengabdian Masyarakat*, *3*(3), 49–60.


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