Issue VI, 22 November 2023
e-ISSN 2707-9481 Institute of Metallurgy and Ore Beneficiation, Satbayev University, Almaty, Kazakhstan
ISBN 978-601-323-356-7 https://doi.org/10.31643/2023.23

Ihda Mutimmatul Fitriyah
Yogyakarta State University (Universitas Negeri Yogyakarta), Jl. Colombo No. 1, Indonesia
E-mail: ihdamutimmatul.2021@student.uny.ac.id

Heri Retnawati
Yogyakarta State University (Universitas Negeri Yogyakarta), Jl. Colombo No. 1, Indonesia
E-mail: heri_retnawati@uny.ac.id
https://orcid.org/0000-0002-1792-5873

Analysis of the distractor of the multiple-choice test using classical test theory (CTT) and item response theory (IRT)

Abstract: This study aimed to analyzed the functional distractor based on the results of the multiple-choice test using CTT and IRT. Data obtained mathematics test at one of the junior high schools (grade 7) in Sidoarjo in the 2021-2022 academic year. The test consists of 20 items, which are a collection of questions that have been standardized in the school curriculum. One hundred examinees attending this mathematics test. The analysis used Iteman 4.3 software for analysis distractor using CTT, and R program for analysis pseudo-guessing's parameter using 3PL IRT model. The result showed that all items have distractors well function it (Prop. > 0,05) in CTT. Meanwhile, with IRT, 8 of the 20 questions were not good because the pseudo-guessing index was > 0.25. The result of this study provided important information for future study to examine the ability estimate when a test's fixed feature is the item-specific characteristic utilized for pseudo-guessing.

Keywords: Distractor, Pseudo-guessing, classical test theory, item response theory.

Cite this article as: Ihda Mutimmatul Fitriyah, Heri Retnawati (2023). Analysis of the distractor of the multiple-choice test using classical test theory (CTT) and item response theory (IRT). *Challenges of Science*. Issue VI, 2023, pp. 196-203. https://doi.org/10.31643/2023.23

Introduction

Multiple choice tests are the most popular test of educational assessment. According to Bolt et al. (2020), multiple-choice tests become a mainstay in an assessment system despite their known limitations. It can happen because multiple-choice tests are easy to score, and offer increased accuracy, reliability, and objectiveness in the assessment process (Becker & Johnston, 1999; Romm et al., 2019; Suseno, 2017; Tangianu et al., 2018; Tarrant & Ware, 2012; Walstad & Becker, 1994; Stepanova et al., 2018). Rodriguez (2016) revealed multiple-choice tests are efficient to administer and take a relatively short time when used for research. Then, Gierl et al. (2017) stated that the most effective, long-lasting, and economical form of assessment is a multiple-choice test.

Multiple choice tests are often used to measure a person's cognitive abilities (Carretta & Ree, 2018; Edwards et al., 2012), both in formative assessment or summative assessment, as well as in school, college, or general such as civil servant selection. Siegfried and Wuttke (2019) also reported that the multiple-choice format is most widely used in college to measure students' cognitive performance. Gierl et al. (2017) revealed that the application of multiple-choice tests for international assessment is PISA and Trends in International Mathematics and Science Study (TIMSS). Therefore, apart from being a popular form of questioning on a national and international scale, multiple-choice tests are also a mainstay in various contexts.

Based on the construct, multiple-choice items consist of a main question (stem) and a set of answer choices (Papenberg & Musch, 2017; Arlinwibowo et al., 2020). In a set of answer choices, there is a correct answer and the others function as distractors, with the number of distractors intended being one or more. A multiple-choice item can be said to be of high quality if the distractors function well (Papenberg & Musch,

2017). Then, Sajjad et al. (2020) stated that at least 5% of good distractors are selected by the total number of examinees.

In the implementation of multiple-choice tests, teachers often ignore the importance of distractors because constructing them is not easy. Gierl et al. (2017) stated that it took a long time to develop distractors. However, there are still many teachers who choose to use multiple-choice tests. The research results by Efrina et al. (2021), found that teachers prefer to create easy distractors and deviate far from the correct answer option with the aim that students will find it easy to answer questions correctly. Apart from that, teachers should avoid creating difficult distractions because it will make the questions difficult for students to solve. The problems above indicate that there needs to be a habit of self-education to be able to carry out multiple-choice tests by paying attention to the function of distractors.

The function of distractors on a multiple-choice item can be estimated by analyzing the characteristics item. Analysis characteristics items can be carried out using two approaches, namely the classical approach (classical test theory) or the item response theory approach. Classical test theory (CTT) is a theory with a simple mathematical model that shows the relationship between observed scores, actual scores, and measurement error (Mardapi, 2012). CTT is applied in estimating reliability, level of difficulty, discrimination index, distractor function, and measurement error (Retnawati, 2017). CTT is considered widely used because it does not require a large number of respondents (more than 100) and is easy to understand and apply (Argianti & Retnawati, 2020). Setiawati et al. (2023) revealed that for more than 20 years, CTT has been the mainstay standard in the development of psychological tests. Even though there are several advantages to CTT, there are things that make it have limitations.

The CTT limitations can be demonstrated in that actual scores are highly dependent on measurements, and test results cannot be compared. Observed scores and actual scores change depending on the level of difficulty and assessment, so both are very dependent on the results of the student's characteristics being measured, where the observed score is the only score that can be seen while the actual score and measurement error are latent (Otaya et al., 2020). Given these limitations, an item response theory (IRT) approach was applied to overcome the limitations of CTT. IRT is widely used in education, research, and psychological measurement practice (Cai et al., 2023). In IRT, the latent trait being measured is called ability (Bahar et al., 2021; Hambleton & Swaminathan, 1985). There must be several equation models involved in the interaction between ability and item parameters. The process of estimating item parameters and abilities can be done directly through the use of the Bayes technique or the maximum likelihood method (Retnawati, 2017).

Item parameters in IRT can be estimated if statistically the model used statisfied the assumptions (Hambleton & Swaminathan, 1985; Santoso et al., 2022). The assumptions are namely unidimensional, local independence, and parameter invariance. This must be fulfilled if an IRT analysis is to be carried out. If the assumptions not statisfied, then the analysis carried out is CTT. Fulfillment of the assumptions is based on the quality of the instrument being tested; therefore, a test developer must have good knowledge so that the questions produced are also good in terms of content quality.

IRT models can be classified in different ways according to the number of response categories. such as Rasch, 1 Parameter Logistic Model (1PL), 2PL, 3PL, and 4PL for dichotomous data; Nominal Response Model (NRM), Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), and Graded Response Model (GRM) for polytomous data (Can Aybek, 2023). Meanwhile, the item parameters resulting from IRT analysis are namely level of difficulty (b), discrimination (a), pseudo-guess (c), and others. Both CTT and IRT analysis can estimate the functioning of distractors; in IRT, it is called pseudo-guessing, where pseudo-guessing's parameter represents the probability of examinees whose abilities are at a low level to be able to answer item i correctly.

The results of research by Huriaty (2016), which analyzed the characteristics of junior high school mathematics tests in the form of multiple-choice questions using IRT 3PL, In the analysis, the Bilog program was assisted, but the 3PL analysis was carried out directly without testing assumptions. The pseudo-guessing estimation results show that the items tested all have good pseudo-guessing indices. Starting from this research, this research will expand and complete previous studies, where the focus of this research is only on distractor parameters whose results want to be synthesized using two approaches, namely CTT and IRT. For IRT, before it is carried out, it is ensured that the assumption test has been met. Meanwhile, for CTT, use the Iteman Program. Thus, this study aims to analyze distractor parameters in multiple-choice tests using CTT and IRT.

Research Methods

***Design and data source*.** This research is a descriptive study with a quantitative approach. The data source is the results of the grade 7 th mathematics test at one of the junior high schools in Sidoarjo in the 2021-2022 academic year. The test consists of 20 items, which are a collection of questions that have been standardized in the school curriculum. School have their own references regarding materials for each semester, they still look at the core competencies-basic competencies (KI-KD) that have been set by the government. These questions consist of 6 items on integers and fractions, 7 items on sets, and 7 items on ratios of two quantities and comparisons. The research was carried out online with the Zoom meetings via Google Form, directly supervised by the mathematic's teacher. A total of 100 students attending this mathematics test.

Data analysis. In accordance with its purpose, this study analyzed the functional distractor based on the results of the mathematics multiple choice test using CTT and IRT. Therefore, data analysis was generally carried out in several stages. First, for CTT, the distractor estimated using Iteman Program (version 4.3), where the results are obtained by looking at the proportion column. Second, the IRT model assumption can fulfill three criteria. They include unidimensional, local independence, and parameter invariance (Retnawati, 2014). Third, estimated the pseudo-guessing parameter (c) with the model fit IRT between 3-PL and 4-PL model. It was used because of its dichotomous scoring, which consists of two categories: the correct answer with a score of 1 and the incorrect answer with a score of 0 (Isnani et al., 2019), also both of that models pseudo-guessing's parameter are estimated. To help analyzed the data, the R program was utilized with the 'mirt' package. The R syntax that we used to estimate under IRT model is available in Appendix 1.

Research Findings and Discussion

***Findings*.** In this section, we report the main findings of this study, namely the distractor of the items based on the estimation using CTT and 3-PL IRT model.

Findings of the Distractor in CTT Approach. In this study, the mathematics multiple choice test used consisted of four options, so that there was one option as the correct answer and three other options as distractors. A distractor is said to function well if it is selected by at least 5% of examinees (see Table 1), and if it does not meet these criteria, then the distractor needs to be revised (Sajjad et al., 2020).

Table 1. Result for The Distractor's Estimation Using Iteman

Item Number	Options			
	A	B	C	D
	Prop.	Prop.	Prop.	Prop.
Item 01	0,1	0,13	0,71*	0,06
Item 02	0,3	0,53*	0,11	0,06
Item 03	0,15	0,45*	0,21	0,19
Item 04	0,15	0,19	0,15	0,51*
Item 05	0,1	0,32*	0,5	0,08
Item 06	0,28	0,2	0,45*	0,07
Item 07	0,1	0,35*	0,47	0,08
Item 08	0,11	0,13	0,66*	0,1
Item 09	0,19	0,14	0,51*	0,16
Item 10	0,09	0,16	0,42	0,33*
Item 11	0,15	0,45*	0,35	0,05
Item 12	0,42*	0,23	0,23	0,12
Item 13	0,28*	0,26	0,18	0,28
Item 14	0,16	0,26	0,45	0,13*
Item 15	0,22*	0,34	0,27	0,17
Item 16	0,1	0,15	0,66*	0,09
Item 17	0,17	0,42	0,24*	0,17
Item 18	0,5*	0,25	0,18	0,07
Item 19	0,09	0,33*	0,4	0,18
Item 20	0,52*	0,21	0,18	0,09

*The correct anwer

Table 1 shows that the distractors from the 20 question items have an answer proportion of 5% to 47%. This is in accordance with the criteria for distractor functioning, namely that distractors are said to function well if they have a minimum proportion of 5%. The distraction, which has a proportion of 5%, is only in one question item, namely item 11.

Findings of the Analysis of IRT Model Assumptions. The first assumptions that has to be statisfied in IRT is the assumption of unidimensional which requires that the mathematics multiple choice test only measure one dominant factor. This can be demonstrated through factor analysis and principal components by considering the eigen value, total variance explained, and the scree plot. The factor analysis can be carried out if qualify the sample adequacy by The Kaiser-Meyer-Olkin (KMO) more than 0,5. The KMO of the mathematics multiple choice test (KMO = 0,540) showed that the sample size has statisfied adequacy for factor analysis. The eigen value for the principal component of 3,038 with the explained variance 15,191%. Scree plot (see Figure 1) shows that there is a steepness from the principal component to two components, for more component it starts to slope, indicates that test is unidimensional.

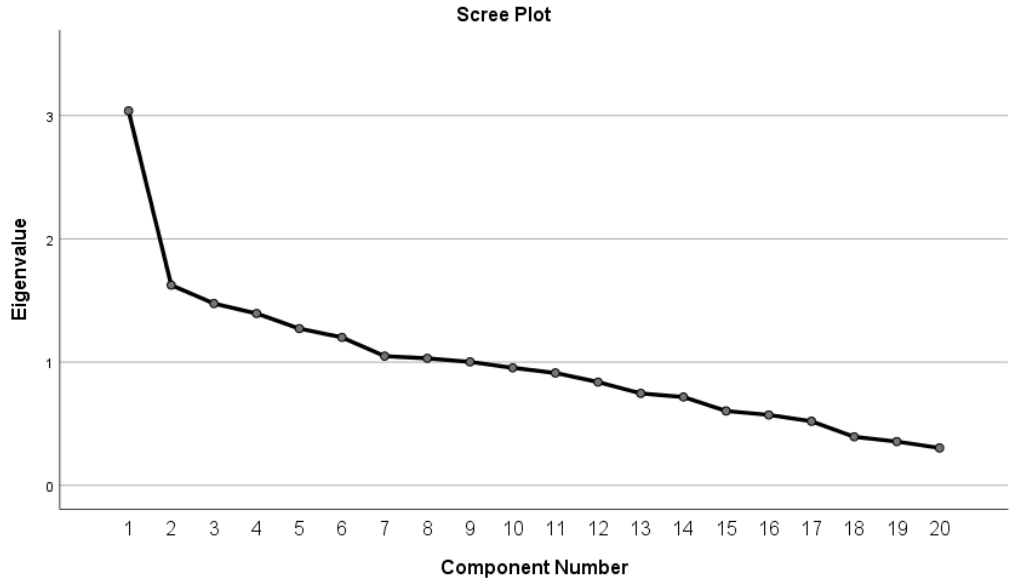

Figure 1. Scree Plot for Unidimensional Assumption

The second assumptions underlying IRT is local independence. This assumption requires that the student's response to an item is independent of his response to other items. Retnawati (2014) stated that local independence automatically fulfilled if the unidimensional assumption is statisfied. Cause the unidimensional assumption test has been proven, so the independence local assumption has also statisfied. The last IRT assumption that we need to show that is parameter invariance. There are two parameters need to prove, namely the item parameter and the person parameter.

The item's parameter invariance is proven by estimating item difficulty for students who take the test by being grouped into two different subgroups based on an even absence and an odd absence. Meanwhile, the person's parameter invariance is proven by estimating student's ability from a subset of items in odd order and a subset of items in even order. The scatter gram showing the distribution of the estimated results for each item's parameter invariance and person's parameter invariance (see Figure 2 and Figure 3). Retnawati (2014) stated that there are dots on a scatter gram (approaches the line that passes through the origin with a gradient of 1). It can be assumed that the item parameters and person parameter are invariance.

Based on Figure 2 and Figure 3, each data has a position relatively close to the line that passes through the origin with a gradient of 1, so that the item parameters and person parameters are invariant. Based on the results above, the three IRT assumptions have been statisfied, so the estimation of item parameters can be continued with 3PL model.

Findings of the Pseudo-guessing's Parameter in IRT Approach. First, determine the fit model between 3PL and 4PL, because both of that model the pseudo-guessing's parameter can be estimated. Basically, model fit can be determined by estimating examinees patterns to the items (Zi Yan & Heene, 2021). Determining the

model fit in this research used Akaike's Information Criterion (AIC), Sample size Adjusted BIC (SABIC), and Bayesian Information Criterion (BIC) values (see Table 2). Data fit the model if these three values are smaller than other IRT models (Djidu et al., 2022).

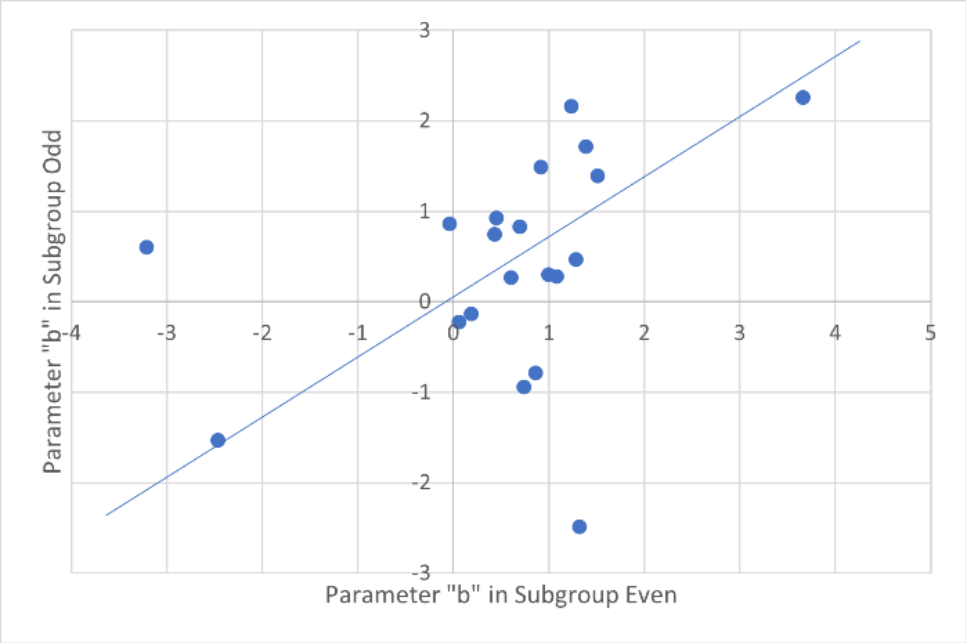

Figure 2. The Scattergram Showing the Distribution of Item Difficulty Estimated from a Subset of Students in Even Absences and Odd Absences

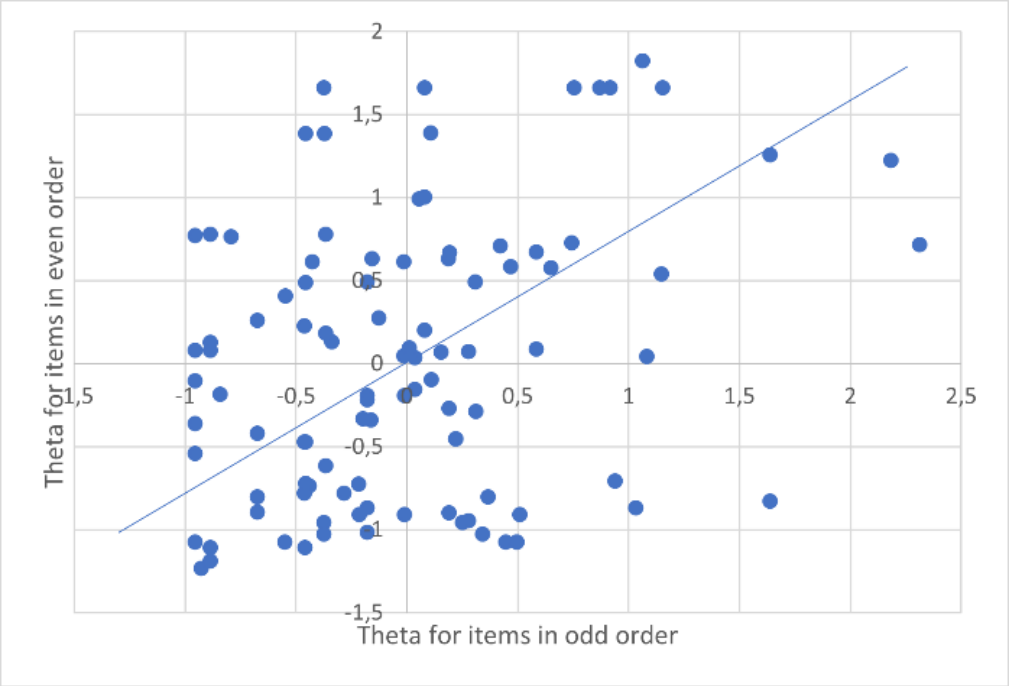

Figure 3. The Scattergram Showing the Distribution of Student's Abilities Estimated from a Subset of Items in Odd Order and Even Order

Table 2. The Model Fit Estimation from The Result of The Mathematics Multiple Choice Test

IRT Model	AIC	SABIC	BIC
3-PL	2520,779	2487,594	2677,089
4-PL	2539,497	2495,250	2747,910

Table 2 shows that the AIC, SABIC, and BIC of 3PL model smaller than 4PL model. It concludes that 3PL model used for further analysis. 3PL model produces three parameters that can be estimated, namely discrimination (a), difficulty (b), and pseudo-guessing (c). Because this research focuses on pseudo-guessing's parameter, so the estimation results shows in Figure 4 and Table 3.

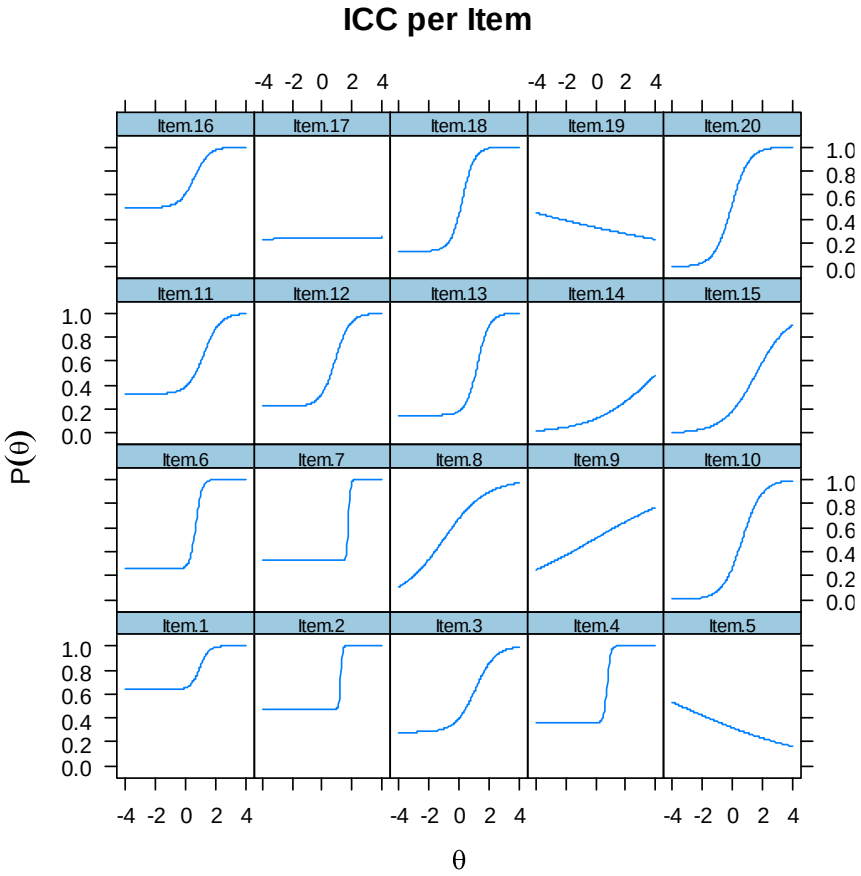

Figure 4. Item Characteristics Curve from The Items

Table 3. The Pseudo-guessing's Parameter Using 3-PL Model Analysis

Item Number	Pseudo-guessing's Parameter	Item Number	Pseudo-guessing's Parameter
Item 01	0,636	Item 11	0,327
Item 02	0,474	Item 12	0,222
Item 03	0,280	Item 13	0,150
Item 04	0,361	Item 14	0,000
Item 05	0,000	Item 15	0,000
Item 06	0,252	Item 16	0,494
Item 07	0,326	Item 17	0,006
Item 08	0,001	Item 18	0,132
Item 09	0,001	Item 19	0,034
Item 10	0,000	Item 20	0,002

Pseudo-guessing's parameter represents the probability of examinees whose abilities are at a low level to be able to answer item i correctly, or the lower asymptote of the ICC of item i. According to Allen & Yen (1979), the pseudo-guessing index is no more than 1/k (*k* being the number of options). Because in this test the number of options is 4, so the apparent pseudo-guessing index is expected to be no more than 0,250. Table 3 shows that the pseudo-guessing index is in the range of 0,000 to 0,634. Eight items (item 01, item 02, item 03, item 04, item 06, item 07, item 11, and item 16) have a pseudo-guessing index of more than 0,250. This indicates that an examinee who has an ability of 0 has a chance of answering correctly (each of the eight items) above 0,250. These results show that the eight items not in good criteria and need to be corrected for the

distractor options. Because the criteria for a good item is that examinees with an ability of 0 should have a chance of guessing the answer correctly with a low value, namely below 0,250, Meanwhile, the other 12 questions already have a pseudo-guessing index of more than 0.250.

Research Discussions

Multiple choice test might be regarded as a popular item types in educational assessment. However, in a test with multiple choice items, some examinees may guess a correct answer (guessing effect). The findings of this study reveal that identifying distractor function can be estimated using two approaches (CTT or IRT). Estimation using IRT can be analyzed if three assumptions have been statisfied, if it is not been statisfied so estimated using CTT. Estimating the effectiveness of distractors with CTT uses the proportion of examinees who answered correctly out of the total examinee for each distractor option. Estimation using this approach is very profitable because it can be identified which distractor options are less functional, so that improvements are only made to the problematic distractor options. This is in line with Fiska et al. (2021) which stated that the effectiveness of the distractor functions in determining the effectiveness of the distractor in carrying out its measuring function and distinguishing between students who understand the concept and those who do not understand the concept. The estimation results show that students do not fully understand the concepts of the mathematics material being tested. In the IRT approach, distractors are termed pseudo-guessing's parameters. Students guess the answer and are correct; in IRT, this is a problem, especially for students with zero (0) ability who should have difficulty guessing the correct option. In contrast to the estimates obtained with CTT, in this IRT, eight of the 20 questions had a pseudo-guessing index below the criterion. This needs to be reviewed again with the existence of these eight questions, so it is indicated that the eight questions have easy difficulty or that distracting options were made at a low level without considering what errors could occur when students choose distracting options. This study was limited to analysis the distractor using two approaches namely CTT and IRT. Furthermore, other item parameters are not estimated, so it cannot be generalized whether the items tested are items with good characteristics or not. Further study, it is recommended to compare the ability estimation when the pseudo-guessing's parameter is item-specific and it is a fixed characteristic of a test. And also, for future research must carry out distractor's analysis on mathematics test with the material is general, so that the result can be generalized.

Conclusion

The findings of this study provide that there is an unfunctional distractors of the mathematics multiple choice test using 3PL IRT model, but for CTT all the distractors have a good function. These findings confirm that between CTT and IRT have a different results estimation but both of these can be used to identify which items must be repaired for get more qualified items of mathematics multiple choice test. For further study, it is recommended to compare the ability estimation when the pseudo-guessing's parameter is item-specific and it is a fixed characteristic of a test. And also, for future research must carry out distractor's analysis on mathematics test with the material is general, so that the result can be generalized.

Cite this article as: Ihda Mutimmatul Fitriyah, Heri Retnawati (2023). Analysis of the distractor of the multiple-choice test using classical test theory (CTT) and item response theory (IRT). *Challenges of Science*. Issue VI, 2023, pp. 196-203. https://doi.org/10.31643/2023.23

References

Argianti, A., & Retnawati, H. (2020). Characteristics of Math national-standardized school exam test items in junior high school: What must be considered? *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(2), 156–165. https://doi.org/10.21831/pep.v24i2.32547

Arlinwibowo J., Kistoro H.C.A., Retnawati H., Kassymova G.K., Kenzhaliyev B.K. (2020). Differences between Indonesia and Singapore based on PISA 2015: Five-factor students' perception in science education. Jurnal Inovasi Pendidikan IPA, 6 (1), pp. 79-87 https://doi.org/10.21831/jipi.v6i1.32637

Bahar, R., Istiyono, E., Widihastuti, W., Munadi, S., Nuryana, Z., & Fajaruddin, S. (2021). Analisis karakteristik soal ujian sekolah hasil musyawarah guru matematika di Tasikmalaya. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *10*(4), 2660. https://doi.org/10.24127/ajpm.v10i4.4359

Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economic understanding. *Economic Record*, *75*(4), 348–357.

Bolt, D. M., Kim, N., Wollack, J., Pan, Y., Eckerly, C., & Sowles, J. (2020). A Psychometric Model for Discrete-Option Multiple-Choice Items. *Applied Psychological Measurement*, *44*(1), 33–48. https://doi.org/10.1177/0146621619835499

Cai, L., Chung, S. W., & Lee, T. (2023). Incremental model fit assessment in the case of categorical data: Tucker–lewis index for item response theory modeling. *Prevention Science*, *24*(3), 455–466. https://doi.org/10.1007/s11121-021-01253-4

Carretta, T. R., & Ree, M. J. (2018). The relations between cognitive ability and personality: Convergent results across measures. *International Journal of Selection and Assessment*, *26*(2–4), 133–144. https://doi.org/10.1111/ijsa.12224

Djidu, H., Ismail, R., RachmaningtyA, N. A., Sumin, Imawan, O. R., Suhariyono, Aviory, K., Prihono, E. W., Kurniawan, D. D., Syahbrudin, J., Nurdin, Marinding, Y., Firmansyah, Retnawati, H., & Hadi, S. (2022). *Analisis Instrumen Penelitian dengan Teori Tes Klasik dan Modern Menggunakan Program R*. UNY Press.

Edwards, B. D., Arthur, W., & Bruce, L. L. (2012). The Three-option Format for Knowledge and Ability Multiple-choice Tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment*, *20*(1), 65–81. https://doi.org/10.1111/j.1468-2389.2012.00580.x

Fiska, J. M., Hidayati, Y., Qomaria, N., & Hadi, W. P. (2021). Analisis butir soal ulangan harian IPA menggunakan software Anates pada pendekatan teori tes klasik. *Natural Science Education Research*, *4*(1), 65–76. https://doi.org/10.21107/nser.v4i1.8133

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research*, *87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. MA: Kluwer Inc.

Hartono, W., Hadi, S., Rosnawati, R., & Retnawati, H. (2022). Uji kecocokan model parameter logistik soal diagnosa kemampuan matematika dasar. *JNPM (Jurnal Nasional Pendidikan Matematika)*, *6*(1), 125–144. https://doi.org/10.33603/jnpm.v6i1.5899

Isnani, I., Utami, W. B., Susongko, P., & Lestiani, H. T. (2019). Estimation of college students' ability on real analysis course using Rasch model. *Research and Evaluation in Education*, *5*(2), 95–102. https://doi.org/10.21831/reid.v5i2.20924

Mardapi, D. (2012). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Nuha Medika.

Otaya, L. G., Kartowagiran, B., Retnawati, H., & Mustakim, S. S. (2020). Estimating the ability of pre-service and in-service Teacher Profession Education (TPE) participants using Item Response Theory. *Research and Evaluation in Education*, *6*(2), 160–173. https://doi.org/10.21831/reid.v6i2.36043

Papenberg, M., & Musch, J. (2017). Of Small Beauties and Large Beasts: The Quality of Distractors on Multiple-Choice Tests Is More Important Than Their Quantity. *Applied Measurement in Education*, *30*(4), 273–286. https://doi.org/10.1080/08957347.2017.1353987

Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika.

Retnawati, H. (2017). Learning trajectory of item response theory course using multiple softwares. *Olympiads in Informatics*, *11*, 123–142. https://doi.org/10.15388/ioi.2017.10

Rodriguez, M. C. (2016). *Selected-response item development. In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development* (2nd ed). Routledge.

Romm, A. T., Schoer, V., & Kika, J. C. (2019). A test taker's gamble: The effect of average grade to date on guessing behaviour in a multiple choice test with a negative marking rule. *South African Journal of Economic and Management Sciences*, *22*(1), 1–13. https://doi.org/10.4102/sajems.v22i1.2542

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

Santoso, A., Pardede, T., Djidu, H., Apino, E., Rafi, I., Rosyada, M. N., & Abd Hamid, H. S. (2022). The effect of scoring correction and model fit on the estimation of ability parameter and person fit on polytomous item response theory. *Research and Evaluation in Education*, *8*(2), 140–151. https://doi.org/10.21831/reid.v8i2.54429

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), 1097–1107. https://doi.org/10.12973/eu-jer.12.2.1097

Siegfried, C., & Wuttke, E. (2019). Are multiple-choice items unfair? And if so, for whom? *Citizenship, Social and Economics Education*, *18*(3), 198–217. https://doi.org/10.1177/2047173419892525

Stepanova G.A., Tashcheva A.I., Stepanova O.P., Menshikov P.V., Kassymova G.K., Arpentieva M.R., Tokar O.V. (2018). The problem of management and implementation of innovative models of network interaction in inclusive education of persons with disabilities. International journal of education and information technologies. Vol. 12, pp. 156-162.

Suseno, I. (2017). Komparasi Karakteristik Butir Tes Pilihan Ganda Ditinjau dari Teori Tes Klasik. *Faktor Jurnal Ilmiah Kependidikan*, *4*(1), 1–8.

Tangianu, F., Mazzone, A., Berti, F., Pinna, G., Bortolotti, I., Colombo, F., Nozzoli, C., Regina, M. La, Greco, A., Filannino, C., Silingardi, M., & Nardi, R. (2018). Are multiple-choice questions a good tool for the assessment of clinical competence in Internal Medicine? *Italian Journal of Medicine*, *12*, 88–96. https://doi.org/10.4081/itjm.2018.980

Tarrant, & Ware. (2012). A framework for improving the quality of multiple-choice assessments. *Nurse Educ*, *37*, 98–104.

Walstad, W., & Becker, W. (1994). Achievement differences on multiple-choice and essay tests in economics. *American Economic Review*, *84*(2), 193–196.

Zi Yan, T. B., & Heene, M. (2021). *Applying The Rasch Model Fundamental Measurement in The Human Sciences* (4 th). Routledge.