Exploring the final examination test item characteristics of Pancasila and civic education

Syarief Fajaruddin [a,1,*], Heri Retnawati [a,2], Eri Yusron [b,3], Vinni Sofyaningsih [c,3]

[a] Universitas Negeri Yogyakarta, Indonesia
[b] Universitas Pendidikan Indonesia, Indonesia
[c] Madrasah Tsanawiyah Negeri 3 Gunungkidul, Indonesia
[1] syarieff@uny.ac.id *; [2] heri_retnawati@uny.ac.id; [3] eyusron98@gmail.com, [4] vinni.soffi@gmail.com

* corresponding author

ARTICLE INFO

Article history
Received 2021-08-24
Revised 2021-09-06
Accepted 2021-12-08
Published 2021-12-29

Keywords
Item response theory
Pancasila and civic education
Final semester examination
2PL
Test item analysis

ABSTRACT

The study aims at exploring the final examination test items to identify the quality of the test items as an evaluation instrument. The study is descriptive quantitative research that pursues the document analysis to view the Final Semester Examination test item characteristics of Pancasila and Civic Education in Grade VIII of State Madrasah Tsanawiyah (Madrasah Tsanawiyah Negeri 3 or MTs Negeri 3) in the Regency of Gunungkidul, the Province of Yogyakarta Special Region, within the 2020/2021 Academic Year. In analyzing the data that have been collected, the researchers have adopted the approach of Item Response Theory (*IRT) using the R Program assistance. The results of the study show that the Final Semester Examination test items of Pancasila and Civic Education are more appropriate to be analyzed by using the SPL Model. The degree of difficulty for these items falls into the "Good" quality within the range -4.0 until +4.0. On the contrary, the item discrimination capacity ranges between 0.079 and 4.891 with the "Moderate" quality.

1. Introduction

In the era of globalization with the advancement of modern technology, education becomes the top priority in order that the development of human resources quality can be the most significant aspect [1]. The reason is that human resources quality depends on education quality [2]–[4]. Well-qualified human resources are the main asset within the development of a nation [5]. In addition, the development of a nation can be viewed from how far the existing education has advanced itself [6]–[8]. In relation to the statement, one of the determiners within the success of education is the high influence by the teacher capacity in implementing the learning activities [9], [10]. Therefore, learning activities are expected to be effective [11], interesting [9], [12], and fun [13]. In addition, another determiner is the necessity for developing various learning models to improve the learning quality [14] and the students' learning results [14]. To identify how far the learning quality and the learning results of the students have progressed, the teachers should conduct an evaluation.

Learning evaluation is one of the ways for attaining information with regards to the overall gain of the students in the aspects of knowledge, concept, attitude, value, and even process skills [15]. Through evaluation, teachers will be able to identify and understand both the individual and the communal achievements of the students [16]. Unfortunately, in Indonesia, there are still a number of issues that have disrupted the learning evaluation. According to Rotama et al. (2020), many teachers still put forward the cognitive aspects of the students and the instruments that teachers design have not undergone any validation process. In addition, the test items are rarely reviewed in terms of validity and reliability, material, construction, language, and even test item analysis after the test items have been administered based on the difficulty level, the item discrimination capacity, and the dummy analysis [17]. This issue appears to the surface due to the limited staff [18] and the busy teaching schedule, leading to the insufficient time for performing the test item analysis and also the lack of knowledge and understanding on the part of the teachers with regards to the test item analysis that should be conducted [17], [19].

The Republic of Indonesia Government Regulation No. 19 of 2005 on National Education Standard mentions that educational assessment on the elementary and high school degree consists of learning results assessment by the educators, learning results assessment by the educational units and learning results assessment by the government [20]. On the contrary, article 64 verse 1 of the same regulation states that the learning results assessment by the teacher as having been intended by article 63 verse 1 is conducted continuously to monitor the process, the progress, and the improvement of the learning results in the form of daily test, mid-semester test, final examination test, and class promotion test. The process of evaluating the students' learning results can be administered by performing the test technique and the non-test technique [18], [21], [22]. Most of the time, teachers implement the test technique in the form of daily tests, mid-semester tests, and final semester tests [18]. All these tests can be either subjective tests or objective tests. In general, the subjective test takes the form of an essay whole the objective test takes the form of a true-false, multiple-choice test, matching test, and completion test [23].

A test can be considered good if the test meets the criteria of validity, reliability, objectivity, practicability, and economics [23], [24]. Thus, the test is expected to provide accurate measurement results. In the context of the study, the test that has been administered is the multiple-choice objective test. According to [25], this type of test offers a number of benefits, namely: (a) being able to measure the learning results objectively; (b) providing faster correction rate; (c) providing faster notification on the scores of the students, and (d) being able to be turned into the test item bank. As an alternative, according to Mania et al. (2020); Slamet dan Maarif (2014), multiple-choice test items offer the following benefit: (a) the test items are easier to analyze; (b) the test items cover many learning materials; (c) all indicators can be met; and (d) students' capacity can be measured in accordance with the desired domain and the difficulty level [26], [27]. Unfortunately, despite those benefits, the multiple-choice test items still suffer from several weaknesses, namely: (a) the designing time is quite demanding; and (b) the designing process takes huge fund resources [25]. Not to mention, the multiple-choice test items are less able to describe the process since they only measure cognitive skills. Therefore, through the multiple-choice test items, the students are able to answer the test items without having to analyze them, and thus, the capacity of the students cannot be completely described [26].

2. Method

The study is descriptive quantitative research that pursues the document analysis to view the characteristics of the Even Semester Examination test item characteristics of Pancasila and Civic Education for Grade VII of MTs Negeri 3 in the Regency of Gunungkidul, the Province of Yogyakarta Special Region, for the 2020/2021 Academic Year (hereinafter shall be referred to MTs Negeri 3 Gunungkidul). The approach adopted within the study is both the qualitative and the quantitative approach, and both approaches have been adopted to view the quality of the final semester examination test item [28]. The subjects within the study are all students from Grade VII of MTs Negeri 3 Gunungkidul with a total number of 161 people. Then, the objects of the study are the responses from the even semester examination test item of Pancasila and Civic Education for Grade VII in the 2020/2021 Academic Year with a total number of 40 items. For the test scoring, the researchers have administered the polytomous scoring with the ordinal scale of 1-2-3-4.

Furthermore, the data-gathering technique that has been implemented is the documentation technique. The documentation technique is implemented to attain the data in the form of even semester examination test item of Pancasila and Civic Education for Grade VII, the answer keys of the test items, and the answer sheets of all Grade VII students from the given subject. The test item analysis using the IRT should meet the assumptions that have been required, and these assumptions are unidimensional assumption, local independent assumption, and invariant parameter assumption. The unidimensional assumption asserts that each test only measures one skill. Thus, the statement implies that every test item only measures one skill of the test takers [29]. In other words, the probability of an item response serves as the single latent characteristic of the test takers [30]. Therefore, a test that has been administered is expected to measure one character or one skill. Then, to meet the unidimensional assumption, the factor that has the most dominant influence on the test performance should be compared to the objective of the test design. If the dominant factor that appears to the surface already meets the objective of the test design, then the unidimensional assumption has already been met. Within the context of the study, the unidimensional assumption testing is conducted by using the SPSS Program.

Next, the local independent assumption defines that the performance of an individual over a test item does not influence the performance of the individual on another test item. This assumption will be met if the response of the test takers on a test item does not influence the response of the test takers on another test item [31]. Last but not least, the invariant parameter assumption defines that the test item characteristics do not depend on the skill parameter distribution of the test takers, and the parameter that becomes the characteristics of the test takers does not depend on the test item characteristics [31]. The implication of this assumption is that the skills of the test takers will not change only because they respond to the test items with different difficult levels [29].

The analysis of the test item is conducted by using the R Program. With regards to the statement, the criteria of item quality within the study refer to the requirements that have been outlined by Hulin et al. (1983), which consists of "Good," "Poor," and "Very Poor." These criteria can be broken down into specific explanations as follows: (a) an item will belong to the "Good" category if the item fits into the model, its difficulty index rangers between -2.0 and 2.0, and its item discriminative index ranges between 0.0 and 2.0; (b) an item will belong to the "Poor" category if the item less fits into the model, its difficulty index ranges is <-2.0 or >2.0, and its item discriminative capacity is >2.0; and (c) an item will belong to the "Very Poor" category if the item does not completely fit into the model

3. Results and Discussion

The results of the study show several information such as item response theory assumption test results, fitness model, item parameter coefficient, item characteristic curve plot, and theta respondent. Within the assumption test, the unidimensional procedure should be conducted first. The unidimensional assumption test is conducted through the exploratory factor analysis using the SPSS program. One of the aspects that should be given attention in performing the exploratory factor analysis is the fulfilment of the sample sufficiency. To identify the sample sufficiency, the values of Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO MSA) can be consulted in Table 1. Based on the analysis results, the KMO value of the instrument is 0.801 with $p < 0.05$. This value is higher than the KMO reference value that has been required, namely ≥ 0.50. In other words, the sample size, 161 respondents, within the analysis has been sufficient. The unidimensional test results are available in Table 1.

The results of factor analysis toward the instrument displayed in Fig 1 show that the instrument within the study only has one dimension. The unidimensional characteristic is apparent since there is only one factor whose eigenvalue has been higher than 1. The eigenvalue of the first factor is 0.871, while the eigenvalue of the second factor is lower than 1.000. in the meantime, the remaining eigenvalues are lower than 1.000. furthermore, within the IRT assumption test, the local independence procedure is also administered. The local independence criteria will be met if the correlation values on each item are lower than 0.200. The local independence test itself is conducted using Yen's Q3. The results of the local independence test show that the highest correlation value is 0.200, while the correlation value of the remaining item is lower than 0.200. Therefore, this finding suggests that the IRT assumption test criteria have been met.

Table 1. KMO and Bartlett's test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.801
Bartlett's Test of Sphericity	Approx. Chi-Square	2614.092
	df	780
	Sig.	.000

Next, the results of the study also display the model fitness. In viewing the model fitness for the analysis, the researchers compare the Akaike Information Criterion (AIC) value, the Bayesian Information Criterion (BYC) value, and the log.Lik value. The lower these values are, the more fit the model in analyzing the data that will be used. The results of the comparison on the model fitness are available in Table 2 and Table 3. Table 2 displays the information on comparing the Rasch Model and the 1PL (1-parameter logistic) Model.

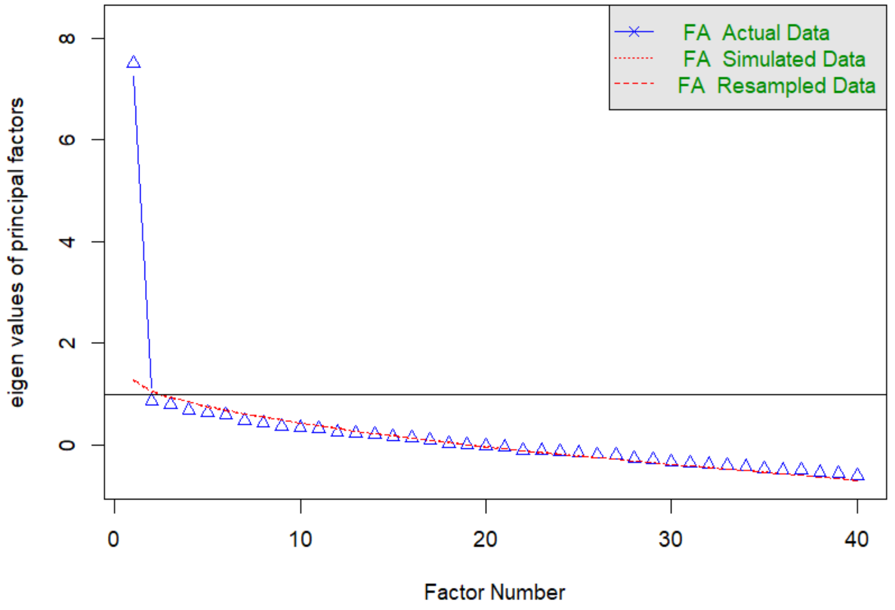

Fig. 1. Unidimensional Test Results

The results of the comparison show that significantly (p.value < 0,001) the AIC value, the BIC value, and the log.Lik values within the 2PL (2-parameter logistic) are lower than the values in the Rasch Model. These findings show that the model fitness comparison between the Rasch Model and the 2PL Model for the data analysis tends to favour the use of the 2 PL Model. Consequently, the researchers should compare the fitness between the 2PL Model and the 3Pl (3-parameter logistic) Model. The results of this comparison are available in Table 4.

Table 2. Likelihood ratio table

	AIC	BIC	log.lik	LRT	df	p.value
Rasch Model	5362.13	5488.46	-2640.06			
2PL Model	5144.82	5391.33	-2492.41	295.31	39	<0.001

Table 3. Likelihood ratio table

	AIC	BIC	log.lik	LRT	df	p.value
Model 2PL	5144.82	5391.33	-2492.41			
Model 3PL	5149.24	5519.01	-2454.62	75.58	40	<0.001

Next, Table 3 shows that the AIC value, the BIC value, and the log.Lik values on the 2PL Model are significantly lower than the values in the 3PL model. Thus, this finding explains that the data within the study are more appropriate to be analyzed using the 2PL Model. Furthermore, Table 4 informs about the difficulty parameter coefficient and the item discriminatory capacity using the 2PL Model a good difficulty index range between -2.00 and +2.00. From the results of the 2PL model analysis toward the test item of Pancasila and civic education final semester examination, it is found that the most difficult item is item number 18, which difficulty coefficient has been 3.422, while the easiest item is item number 27, which difficulty coefficient has been 3.076. The difficulty degree of the 40 test items itself ranges between -4.00 and +4.00. In general, the item difficulty falls into the coefficient -0.979 with the difficulty standard deviation of 1.139.

Table 4. Item Parameter Estimation

	Difficulty	Category	Discriminatory	Category
Item1	-0.1488974	Good	0.84275811	Good
Item2	-0.7082914	Good	1.53283789	Good
Item3	-1.4431486	Good	1.76268559	Good
Item4	-0.9928439	Good	1.37829961	Good
Item5	-1.1267938	Good	0.88592059	Good
Item6	-0.8301167	Good	1.26470241	Good
Item7	-1.6555222	Good	1.89100836	Good
Item8	-1.7850678	Good	1.35617743	Good
Item9	-0.9043582	Good	2.77097585	Poor
Item10	2.3757943	Poor	0.45176132	Good
Item11	-1.8916202	Good	1.49214763	Good
Item12	1.2803417	Good	0.41675797	Good
Item13	-1.8492907	Good	0.98101151	Good
Item14	-1.5751570	Good	1.00505962	Good
Item15	-1.7904338	Good	2.09008074	Poor
Item16	-1.3904576	Good	0.83969626	Good
Item17	-0.9815546	Good	3.79559935	Poor
Item18	3.4216599	Poor	-0.07950051	Good
Item19	-1.2354269	Good	2.04031372	Poor
Item20	-1.4459365	Good	1.38560385	Good
Item21	-0.9168343	Good	0.82343146	Good
Item22	-2.2123674	Poor	1.35908205	Good
Item23	-1.7099337	Good	3.65244094	Poor
Item24	-0.8778555	Good	1.16804883	Good
Item25	-0.2314381	Good	1.20772821	Good
Item26	-1.0825430	Good	2.89805551	Poor
Item27	-3.0759335	Poor	-0.75497190	Good
Item28	-0.7117559	Good	1.99371958	Good
Item29	-0.9085943	Good	2.02857020	Poor
Item30	-1.2139848	Good	4.89110889	Poor
Item31	-0.8155556	Good	4.20709309	Poor
Item32	-1.1933874	Good	1.84579498	Good
Item33	-1.5130661	Good	1.89725576	Good
Item34	-1.0917900	Good	3.09343795	Poor
Item35	-1.5986372	Good	3.53536401	Poor
Item36	-0.6370078	Good	2.38795452	Poor
Item37	-0.9240531	Good	1.95627267	Good
Item38	-2.1969043	Poor	-1.44694226	Good
Item39	-0.7689280	Good	2.94109680	Poor
Item40	-0.8221785	Good	1.21640791	Good

This finding shows that overall, the instrument has a good difficulty level in each item. In relation to the statement, the 2PL logistic model analysis yields different discriminatory capacities for each item. In this regard, the results in Table 4 show the discriminatory capacity range between 0.0079 and 4.891. The item with the lowest discriminatory capacity is item number 18, which index is 0.079, while the item with the highest discriminatory capacity is item number 30, which index is 4.891. There are several ICCs (Item Characteristic Curve) from several items, and these curves are available in Fig 2.

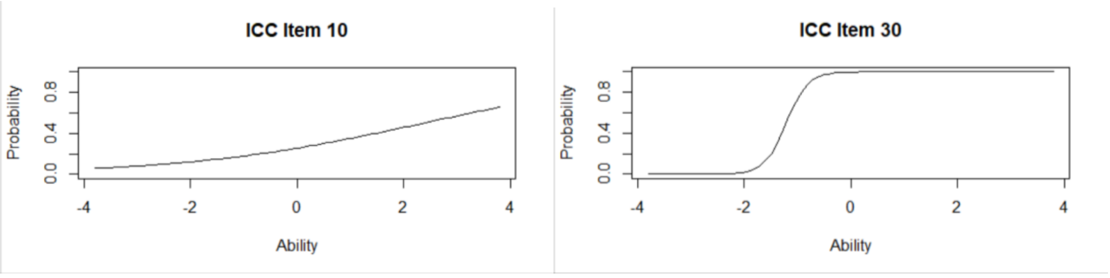

Fig. 2. The ICC for the Item Number 10 and Number 30

Fig 2 refers to the ICC for the item number and item number 30. The ICC for item number 10 describes that the item has a high difficulty index, as is apparent from the very slopy curve line.

Therefore, it is apparent that the probability for the respondents with theta 2 to respond to the test item correctly is only 0.40. Thus, the test item is moderate in discriminating the respondent capacity. In the meantime, the ICC for test item number 30 shows that the curve line falls between -2.00 and -1.00, which is very steep. This finding shows that the item discriminatory capacity for the test item number 30 is poor and, thus, it has a low difficulty index.

Table 5. Test Participant Capacity

	EAP	EB	MI
Mean	-0.0803	-0.1441	-0.1287
SD	0.8232	0.7800	0.7547
Max.	1.6915	1.5617	1.5604
Min.	-2.0702	-2.0373	-1.9980

The item analysis using the IRT should meet the three assumptions that have been required [30]. The assumptions that have been generally used in the IRT models are the unidimensional assumption, the local independent assumption, and the invariant parameter assumption [31], [33], [34]. The unidimensional test is conducted in order to identify whether a test measures only one trait or not [30]. In this regard, the results of the analysis show that the instrument has been confirmed to have only one dimension, namely measuring the student capacity in Pancasila and Civic Education. Then, in the local independent test, the results of the analysis show that the highest correlation value is 0.200 while the remaining correlation value is lower than 0.200. Thus, this finding implies that the IRT assumption test within the instrument has already been met. This finding is in accordance with the results of a study by Hambleton dan Swaminathan (1985), which state that if the covariant value of the students' skills group is closer to zero, then the local independent assumption test criteria have already been met. In addition, the local independent assumption test will be met if the test has been confirmed to be unidimensional [35]–[37]. The local independent test is performed to identify the students' response on a test item and the response should not be dependent on their response to the other item.

The fitness model that has been adopted in the study is the 2PL Model. This finding is attained after the instrument is compared in terms of compatibility among the 1PL (Rasch) Model, the 2PL Model, and the 3PL Model. The lower the values in the data analysis are, the fitter, the model, will be in analyzing the data that will be used within the study. The statement is in line with the results of a study by Jafar, which state that the parameter model that shows the lowest Akaike Information Criterion value is a fit for use. The results of the test item analysis for the Final Semester Examination of Pancasila and Civic Education in Grade VII MTs Negeri 3 Gunungkidul using the 2PL Model inform about the difficulty level (bi) and the discriminatory capacity (ai). The results of the analysis show that the instrument, overall, has a good difficulty index for each item. These results have been confirmed with the 35 test items (87.50%) that belong to the "Good" category and 5 test items (12.50%) that belong to the "Poor" category. This conclusion is based on the range between -2.00 and +2.00 within the logit scale [34], [35], [39], [40]. In the meantime, the discriminatory capacity information within the instrument shows sufficient results. The statement is based on the argument by DeMars (2018), who states that the good discriminatory capacity range between 0.00 and +2.00. Therefore, it can be safely concluded that the test item instrument of Pancasila and Civic Education in MTs Negeri 3 Gunungkidul is able to differentiate between the high-performing students and the low-performing students.

4. Conclusions

The test item analysis for the Final Semester Examination of Pancasila and Civic Education in MTs Negeri 3 Gunungkidul shows that the test item is more appropriate to be analyzed by using the 2PL Model. Based on the analysis using the 2PL Model, it is found that the most difficult item is test item number 18 while the easiest test item is item number 27, with the degree of difficulty that falls into the range between -4.00 and +4.00. In general, the coefficient of the difficulty index is -0.979, with the difficulty standard deviation of 1.139. Thus, this finding implies that the instrument has a good difficulty index. In the meantime, the item discriminatory capacity falls into the range between 0.079 and 4.891. The item with the lowest discriminatory index is item number 18, while the item with the highest discriminatory index is item number 30.

Acknowledgment

The researchers would like to thank the Lecturer of Item Response Theory for the guidance and all the colleagues from PEP Doctoral Program Batch 2020 for the support. Then, the researchers would also like to thank Jean Antunes Rudolf Zico Ma'u for the willingness to translate the article.

Declarations

Author contribution	:	SF: Concepting and composing the manuscript; HR: Guiding and directing; EY: Processing the data; VS: Collecting the data
Funding statement	:	The author received no financial support for the research and/or authorship of this article
Conflict of interest	:	The author declares no conflict of interest.
Additional information	:	No additional information is available for this paper.

References

[1] F. Malia and J. Simanjuntak, "Jokowi: Pendidikan perlu revolusi mental," *Tribun News*, 2014. .

[2] K. Nurhadi, *Kurikulum 2004: Pertanyaan dan jawaban*. Jakarta: Grasindo, 2004, available at: Google Scholar.

[3] L. Octavianus, "Perbedaan sikap guru terhadap program 5 hari belajar efektif ditinjau dari latar belakang pendidikan pada guru-guru di SMU Negeri 10 Medan," Universitas Medan Area, 2008, available at: Google Scholar.

[4] D. Handayani, "Peningkatan pemahaman siswa tentang kenampakan alam melalui metode diskusi pada siswa kelas IV Semester 1 SDN 1 Bicak Todanan-Blora Tahun 2015/2016," Universitas Muhammadiyah Surakarta, 2015, available at: Google Scholar.

[5] M. Muhardi, "Kontribusi pendidikan dalam meningkatkan kualitas bangsa Indonesia," *Mimb. J. Sos. dan Pembang.*, vol. 20, no. 4, pp. 478–492, 2004, doi: 10.29313/mimbar.v20i4.153.

[6] D. Marnasari, "Pengaruh Konsep Diri Dan Kebiasaan Belajar Terhadap Prestasi Belajar Akuntansi Siswa Jurusan Akuntansi Kelas XI SMK Muhammadiyah 2 Klaten Utara Tahun Ajaran 2010/2011," Universitas Muhammadiyah Surakrta, 2011, available at: Google Scholar.

[7] A. Permatasari, "Membangun kualitas bangsa dengan budaya literasi," 2015, available at: Google Scholar.

[8] R. Hapipah, "Pengaruh kurangnya fasilitas belajar mengajar untuk siswa dalam mengembangkan pendidikan," Universitas Lambung Mangkurat, 2021. doi: 10.5281/zenodo.4893250.

[9] A. D. Rotama, T. W. Budiutomo, and A. N. A. Bowo, "Analisis butir soal penilaian tengah semester mata pelajaran PPKN kelas VII di SMP Muhammadiyah 7 Yogyakarta," *Acad. Educ. J.*, vol. 11, no. 01, pp. 24–35, Jan. 2020, doi: 10.47200/aoej.v11i01.314.

[10] A. N. A. Bowo, *Cerita cinta belajar mengajar*. Deepublish, 2015, available at: Google Scholar.

[11] P. Setyosari, "Menciptakan pembelajaran yang efektif dan berkualitas," *JINOTEP (Jurnal Inov. dan Teknol. Pembelajaran) Kaji. dan Ris. dalam Teknol. Pembelajaran*, vol. 1, no. 1, pp. 20–30, Dec. 2017, doi: 10.17977/um031v1i12014p020.

[12] D. S. Purnama, "Implementasi model pembelajaran kreatif dan produktif dalam upaya peningkatan mutu pendidikan guru," *Maj. Ilm. Pembelajaran*, vol. 4, no. 2, 2008, available at: Google Scholar.

[13] A. Amir, "Penggunaan media gambar dalam pembelajaran matematika," *Eksakta J. Penelit. dan Pembelajaran MIPA*, vol. 2, no. 1, pp. 34–40, 2016, doi: 10.31604/eksakta.v1i2.%25p.

[14] T. Nurrita, "Pengembangan media pembelajaran untuk meningkatkan hasil belajar siswa," *MISYKAT J. Ilmu-ilmu Al-Quran, Hadist, Syari'ah dan Tarb.*, vol. 3, no. 1, pp. 171–210, 2018, doi: 10.33511/misykat.v3n1.171.

[15] N. Nuriyah, "Evaluasi pembelajaran: sebuah kajian teori," *Edueksos J. Pendidik. Sos. Ekon.*, vol. 3, no. 1, 2016, available at: Google Scholar.

[16] N. S. Raharja, “Analisis butir soal ujian akhir sekolah produktif pemasaran kelas XII Pemasaran SMK Negeri 9 Semarang,” *Econ. Educ. Anal. J.*, vol. 3, no. 3, 2014, available at: Google Scholar.

[17] M. Afandi and I. I. Nazilah, “Analisis soal UAS gasal 2017/2018 PKn kelas VI SD di UPTD Pendidikan Genuk Semarang,” *Tunjuk Ajar J. Penelit. Ilmu Pendidik.*, vol. 1, no. 2, p. 103, Aug. 2018, doi: 10.31258/jta.v1i2.103-115.

[18] W. S. Oktanin and S. Sukirno, “Analisis butir soal ujian akhir mata pelajaran ekonomi akuntansi,” *J. Pendidik. Akunt. Indones.*, vol. 13, no. 1, Jun. 2015, doi: 10.21831/jpai.v13i1.5183.

[19] I. I. Nazilah, “Analisis soal ulangan akhir semester gasal tahun pelajaran 2017/2018 mata pelajaran PKn kelas VI SD di Kecamatan Genuk Semarang.” Universitas Islam Sultan Agung, 2018, available at: Google Scholar.

[20] Presiden Republik Indonesia, *Peraturan Pemerintah Republik Indonesia no 19 th 2005 tentang Standar Nasional Pendidikan*. Indonesia, 2005, available at: Google Scholar.

[21] I. Hakim, “Pengembangan alat evaluasi hasil belajar siswa (teknik tes) berbasis permainan (game) edukatif mata pelajaran ekonomi kelas X IIS Semester Genap SMA Negeri 6 Malang,” Universitas Negeri Malang, 2016, available at: Google Scholar.

[22] S. Sawaluddin and S. Muhammad, “Langkah-langkah dan teknik evaluasi hasil belajar Pendidikan Agama Islam,” *J. PTK dan Pendidik.*, vol. 6, no. 1, Jul. 2020, doi: 10.18592/ptk.v6i1.3793.

[23] S. Arikunto, *Dasar-dasar evaluasi pendidikan*, 3th ed. Jakarta: Bumi Aksara, 2016, available: Google Scholar.

[24] H. Supiyansyah, I. H. Kusumah, and E. T. Berman, “Analisis kualitas soal ulangan akhir semester genap pada mata pelajaran produktif program keahlian teknik kendaraan ringan,” *J. Mech. Eng. Educ.*, vol. 4, no. 1, pp. 52–58, 2017, doi: 10.17509/jmee.v4i1.7441.

[25] E. A. Wibawa, “Karakteristik butir soal tes ujian akhir semester hukum bisnis,” *J. Pendidik. Akunt. Indones.*, vol. 17, no. 1, pp. 86–96, 2019, doi: 10.21831/jpai.v17i1.26339.

[26] S. Mania, F. Fitriani, A. F. Majid, N. N. Ichiana, and A. I. P. Abrar, “Analisis butir soal ujian akhir sekolah,” *Al asma J. Islam. Educ.*, vol. 2, no. 2, p. 274, Nov. 2020, doi: 10.24252/asma.v2i2.16569.

[27] S. Slamet and S. Maarif, “Pengaruh bentuk tes formatif assosiasi pilihan ganda dengan reward dan punishment score pada pembelajaran matematika siswa SMA,” *Infin. J.*, vol. 3, no. 1, p. 59, Feb. 2014, doi: 10.22460/infinity.v3i1.39.

[28] D. D. Kurniawan, “Analisis kualitas soal ujian akhir semester matematika berdasarkan teori respon butir,” 2015, available at: Google Scholar.

[29] A. A. P. Antara, *Penyetaraan vertikal dengan pendekatan klasik dan item response theory (teori dan aplikasi)*. Deepublish, 2020, available at: Google Scholar.

[30] F. Friyatmi, “Estimasi parameter tes dengan penskoran politomus menggunakan graded response model pada sampel kecil,” *J. Inov. Pendidik. Ekon.*, vol. 8, no. 1, pp. 22–31, 2018, doi: 10.24036/01104490.

[31] H. Retnawati, *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Yogyakarta: Nuha Medika, 2014, available at: Google Scholar.

[32] C. L. Hulin, F. Drasgow, and C. K. Parsons, *Item response theory: Application to psychological measurement*. Homewood, Illinois: Dow Jones-Irwin, 1983, available at: Google Scholar.

[33] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of item response theory*. Sage, 1991, available at: Google Scholar.

[34] R. K. Hambleton and H. Swaminathan, *Item response theory*. Dordrecht: Springer Netherlands, 1985, available at: Google Scholar.

[35] C. E. DeMars, “Classical test theory and item response theory,” in *The Wiley Handbook of Psychometric Testing*, Chichester, UK: John Wiley & Sons, Ltd, 2018, pp. 49–73, doi: 10.1002/9781118489772.ch2.

[36] A. Santoso, K. Kartianom, and G. K. Kassymova, “Kualitas butir bank soal statistika (Studi kasus: Instrumen ujian akhir mata kuliah statistika Universitas Terbuka),” *J. Ris. Pendidik. Mat.*, vol. 6, no. 2, pp. 165–176, 2019, doi: 10.21831/jrpm.v6i2.28900.

[37] R. Danni, A. Wahyuni, and T. Tauratiya, “Item response theory approach: Kalibrasi butir soal penilaian akhir semester mata pelajaran bahasa Arab,” *Arab. J. Arab. Stud.*, vol. 6, no. 1, pp. 93–104, 2021, doi: 10.24865/ajas.v6i1.320.

[38] R. A. Jafar, M. Mansyur, and T. Pristiwaluyo, “Konsistensi peserta dalam menjawab soal USBN IPA tingkat SD dengan menggunakan metode Person Fit,” Universitas Negeri Makassar, 2019, available at: Google Scholar.

[39] B. Sumintono and W. Widhiarso, *Aplikasi pemodelan Rasch pada assessment pendidikan*. Cimahi: Trim Komunikata, 2015, available at: Google Scholar.

[40] D. Mardapi, “Analisis butir dengan teori tes klasik dan teori respons butir,” *J. Kependidikan*, vol. 28, no. 2, 1998, doi: 10.21831/jk.v28i1.7244.