Proceedings Series on Social Sciences & Humanities, Volume 12
Proceedings of International Conference on Social Science (ICONESS)
ISSN: 2808-103X

Analysis of Final Summative Assessment Test Items for Grade 6 Elementary School in Physical Education Using Anates

Deppy Widianto[1], Wanda Nugroho Yanuarto[2], Subuh Anggoro[3]
[1]SD Negeri Kutasari 03 Cipari, Cilacap
[2,3]Magister Pendidikan Dasar, Universitas Muhammadiyah Purwokerto

ARTICLE INFO

Article history:

DOI:
10.30595/pssh.v25i.1695

Submited:
July 22, 2025

Accepted:
August 11, 2025

Published:
August 24, 2025

Keywords:

Item Analysis; ANATES; Physical Education; Reliability; Difficulty Level; Discriminating Power; Distractor Effectiveness

ABSTRACT

This study evaluated the quality of test items in the Physical Education, Sports, and Health (PJOK) Final Summative Assessment (ASAT) for Grade 6 students at SDN Kutasari 03 Cipari, Cilacap Regency. Using ANATES software, 25 multiple-choice and 10 essay questions were analyzed from a sample of 5 students. The analysis focused on reliability, discriminating power, difficulty level, and distractor effectiveness. Results revealed high reliability for multiple-choice questions (0.72) and very high reliability for essay questions (0.93). For multiple-choice items, 48% demonstrated good discriminating power, 32% had balanced difficulty levels, and 40% showed significant correlations with total scores. Among essay questions, 60% exhibited good discriminating power, and all items had moderate difficulty levels. Based on these findings, several test items require revision to enhance assessment quality in the Physical Education subject, particularly focusing on improving discriminating power and difficulty balance for optimal student evaluation..

Corresponding Author:
Deppy Widianto
SD Negeri Kutasari 03 Cipari, Cilacap
Email: dhepare@gmail.com

1. INTRODUCTION

Learning outcome evaluation is an important component in the learning process. Through evaluation, teachers can measure the extent to which students have achieved the established learning objectives. Additionally, evaluation also functions as feedback for teachers to improve the quality of learning [1].

In the context of formal education, learning outcome evaluation is generally conducted through written tests, either in the form of objective tests (multiple choice) or subjective tests (essays). For tests to accurately measure students' abilities, the test instruments used must meet certain criteria, such as validity, reliability, objectivity, and practicability [2].

Physical Education, Sports, and Health (PJOK) is one of the mandatory subjects at the elementary school level. This subject aims to develop the physical, mental, social, and emotional aspects of students through physical activities [3]. Although PJOK involves more practical activities, cognitive evaluation is still necessary to measure students' understanding of the basic concepts of physical education and health.

Item analysis is the process of examining the quality of test items, both qualitatively and quantitatively. Quantitative analysis is conducted to determine the statistical parameters of test items, such as difficulty level, discriminating power, and distractor effectiveness (for multiple-choice questions). Through item analysis, teachers can identify good items to retain, items that need revision, and items that should be replaced [4].

ANATES (Test Analysis) is one software that can be used to conduct practical and quick item analysis. This program was developed by Drs. Karno To, M.Pd. and Yudi Wibisono, ST. from the Indonesian University of Education (UPI). ANATES can be used to analyze both multiple-choice and essay test items [5].

Based on this background, this research aims to analyze the quality of test items in the Final Summative Assessment (ASAT) for 6th Grade PJOK subjects at SDN Kutasari 03 Cipari, Cilacap Regency using ANATES software.

2. RESEARCH METHOD

This research is a descriptive study with a quantitative approach. Descriptive research aims to describe or illustrate existing phenomena, both natural phenomena and human-engineered phenomena [6]. The subjects of this study were 5 Grade 6 students at SDN Kutasari 03 Cipari, Cilacap Regency who took the Final Summative Assessment (ASAT) for PJOK. The object of this research is the test items of the Final Summative Assessment (ASAT) for Grade 6 subjects, consisting of 25 multiple-choice items and 10 essay items.

Data in this study were collected through the documentation method, by collecting documents in the form of test sheets for the Final Summative Assessment (ASAT) for Grade 6 subjects and student answer sheets. Data were analyzed using ANATES version 4 program. This program can analyze both multiple-choice and essay test items [7]. The parameters analyzed include:

a. Test reliability

Reliability is the level of consistency or stability of a test instrument in measuring what it should measure. Reliability shows the extent to which measurement results can be trusted. The reliability coefficient ranges from 0 to 1. The higher the reliability coefficient, the higher the level of confidence in the measurement results.

b. Discriminating power

Discriminating power is the ability of a test item to differentiate between high-ability test takers (upper group) and low-ability test takers (lower group). Discriminating power is expressed in the form of a discrimination index (D) ranging from -1.00 to 1.00.

c. Difficulty level

Difficulty level is the probability of answering a test item correctly at a certain ability level. Difficulty level is expressed in the form of a difficulty index (p) ranging from 0.00 to 1.00.

d. Distractor effectiveness (for multiple-choice questions)

Distractor effectiveness is a measure of how well incorrect answer choices can mislead test takers who do not know the correct answer. A distractor is said to function well if it is chosen by at least 5% of the total number of test takers.

e. Correlation of item scores with total scores

Correlation of item scores with total scores is a statistical analysis that measures the relationship or connection between scores obtained by students on each test item and the overall total test score. This is one method to determine the validity of test items in evaluation instruments.

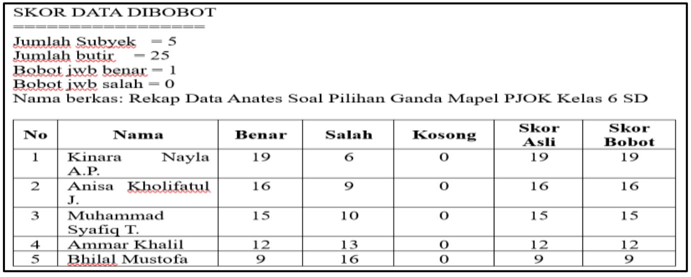
The analysis results are then interpreted based on established criteria for each parameter.

3. RESULTS AND DISCUSSIONS

3.1 Research Results

3.1.1. Multiple-Choice Question Analysis

ANATES Data Recap for Multiple-Choice Questions. Figure 1 will explain the recap of multiple-choice questions for Grade 6 PJOK subjects using ANATES Application version 4.



```
SKOR DATA DIBOBOT
====================
Jumlah Subyek   = 5
Jumlah butir    = 25
Bobot jwb benar = 1
Bobot jwb salah = 0
Nama berkas: Rekap Data Anates Soal Pilihan Ganda Mapel PJOK Kelas 6 SD
```

No	Nama	Benar	Salah	Kosong	Skor Asli	Skor Bobot
1	Kinara Nayla A.P.	19	6	0	19	19
2	Anisa Kholifatul J.	16	9	0	16	16
3	Muhammad Syafiq T.	15	10	0	15	15
4	Ammar Khalil	12	13	0	12	12
5	Bhilal Mustofa	9	16	0	9	9

Figure 1. ANATES Data Recap for Multiple-Choice Questions

Test Reliability. The analysis results show that the reliability coefficient for multiple-choice tests is 0.72. Based on criteria proposed by Jannah [8], the reliability of multiple-choice tests falls into the high category.

Discriminating Power. It can be seen that from 25 multiple-choice items, 12 items (48%) have very good discriminating power and 13 items (52%) have poor discriminating power. This will be explained in Table 1 as follows.

Table 1 Distribution of Multiple-Choice Items Based on Discriminating Power

Category	Discriminating Power Index	Number of Items	Percentage	Item Numbers
Very Good	0,71 - 1,00	12	48%	1, 3, 4, 6, 7, 10, 15, 16, 17, 18
Good	0,41 - 0,70	0	0%	-
Fair	0,21 - 0,40	0	0%	-
Poor	0,00 - 0,20	13	52%	2, 5, 8, 9, 11, 12, 13, 14, 19, 20, 21, 22, 23, 24, 25
Total		25	100%	

Difficulty Level. It can be seen that from 25 multiple-choice items, 1 item (4%) falls into the very difficult category, 7 items (28%) fall into the difficult category, 8 items (32%) fall into the moderate category, 3 items (12%) fall into the easy category, and 6 items (24%) fall into the very easy category. This will be explained in Table 2 as follows.

Table 2. Distribution of Multiple-Choice Items Based on Difficulty Level

Category	Difficulty Index	Number of Items	Percentage	Item Numbers
Very Difficult	0,00 - 0,19	1	4%	13
Difficult	0,20 - 0,30	7	28%	4, 5, 7, 10, 16, 20, 25
Moderate	0,31 - 0,70	8	32%	1, 3, 6, 9, 15, 19, 23
Easy	0,71 - 0,80	3	12%	11, 17, 18
Very Easy	0,81 - 1,00	6	24%	2, 8, 12, 14, 21, 22, 24
Total		25	100%	

Correlation of Item Scores with Total Scores. It can be seen that from 25 multiple-choice items, 10 items (40%) have a very significant correlation with the total score, 10 items (40%) have a non-significant correlation with the total score, and 5 items (20%) cannot have their correlation calculated (NAN). This will be explained in Table 3 as follows.

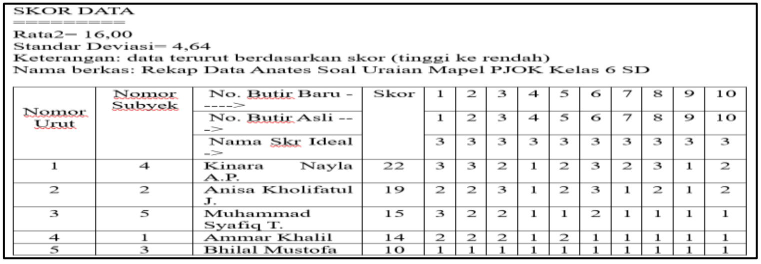
Table 3. Distribution of Multiple-Choice Items Based on Correlation of Item Scores with Total Scores

Category	Correlation	Number of Items	Percentage	Item Numbers
Very Significant	> 0,576	10	40%	1, 3, 4, 6, 7, 10, 15, 16, 17, 18
Not Significant	< 0,576	10	40%	5, 9, 11, 19, 20, 23, 25
Cannot Be Calculated	NAN	5	20%	2, 8, 12, 13, 14, 21, 22, 24
Total		25	100%	

Distractor Effectiveness. The analysis results show that from 75 distractors (3 distractors on 25 items), 19 distractors (25.33%) function very well, 22 distractors (29.33%) function well, 3 distractors (4%) function fairly, 16 distractors (21.33%) function poorly, and 15 distractors (20%) function very poorly.

3.1.2. Essay Question Analysis

ANATES Data Recap for Essay Questions. Figure 2 will explain the recap of essay questions for Grade 6 PJOK subjects using ANATES Application version 4.



SKOR DATA
==========
Rata2= 16,00
Standar Deviasi= 4,64
Keterangan: data terurut berdasarkan skor (tinggi ke rendah)
Nama berkas: Rekap Data Anates Soal Uraian Mapel PJOK Kelas 6 SD

Nomor Urut	Nomor Subyek	No. Butir Baru ----->	Skor	1	2	3	4	5	6	7	8	9	10
		No. Butir Asli --->		1	2	3	4	5	6	7	8	9	10
		Nama Skr Ideal -->		3	3	3	3	3	3	3	3	3	3
1	4	Kinara Nayla A.P.	22	3	3	2	1	2	3	2	3	1	2
2	2	Anisa Kholifatul J.	19	2	2	3	1	2	3	1	2	1	2
3	5	Muhammad Syafiq T.	15	3	2	2	1	1	2	1	1	1	1
4	1	Ammar Khalil	14	2	2	2	1	2	1	1	1	1	1
5	3	Bhilal Mustofa	10	1	1	1	1	1	1	1	1	1	1

Figure 2. ANATES Data Recap for Essay Questions

Test Reliability. The analysis results show that the reliability coefficient for essay tests is 0.93. Based on criteria proposed by Jannah [8], the reliability of essay tests falls into the very high category.

Discriminating Power. It can be seen that from 10 essay items, no items have very good discriminating power, 6 items (60%) have good discriminating power, 2 items (20%) have fair discriminating power, and 2 items (20%) have poor discriminating power. This will be explained in Table 4 as follows.

Table 4. Distribution of Essay Items Based on Discriminating Power

Category	Discriminating Power Index	Number of Items	Percentage	Item Numbers
Very Good	0,71 - 1,00	0	0%	-
Good	0,41 - 0,70	6	60%	1, 2, 6, 8
Fair	0,21 - 0,40	2	20%	3, 5, 7, 10
Poor	0,00 - 0,20	2	20%	4, 9
Total		10	100%	

Difficulty Level. It can be seen that from 10 essay items, all items (100%) fall into the moderate category. This will be explained in Table 5.

Table 5. Distribution of Essay Items Based on Difficulty Level

Category	Difficulty Index	Number of Items	Percentage	Item Numbers
Difficult	0,00 - 0,30	0	0%	-
Moderate	0,31 - 0,70	10	100%	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Easy	0,71 - 1,00	0	0%	-
Total		10	100%	

Correlation of Item Scores with Total Scores. It can be seen that from 10 essay items, 8 items (80%) have a very significant correlation with the total score and 2 items (20%) cannot have their correlation calculated (NAN). This will be explained in Table 6 as follows.

Table 6. Distribution of Essay Items Based on Correlation of Item Scores with Total Scores

Category	Correlation	Number of Items	Percentage	Item Numbers
Very Significant	> 0,576	8	80%	1, 2, 3, 5, 6, 7, 8, 10
Cannot Be Calculated	NAN	2	20%	4, 9
Total		10	100%	

3.1.3. Quality of Multiple-Choice Questions

The analysis results show that the reliability coefficient of the multiple-choice test is 0.72, which falls into the high category. This indicates that the multiple-choice test has high consistency in measuring students' abilities. However, considering that test reliability is influenced by the number of items and the number of test participants, a reliability of 0.72 with 25 items and 5 participants is quite good. This is in line with Sukiman's [9] opinion that the minimum reliability coefficient for teacher-made tests is 0.70. Based on the analysis results, 12 items (48%) have very good discriminating power and 13 items (52%) have poor discriminating power. Items with very good discriminating power show that these items can differentiate between high-ability test takers and low-ability test takers.

Conversely, items with poor discriminating power show that these items cannot differentiate between high-ability test takers and low-ability test takers. Arifin and Retnawati [10] suggest that items with poor discriminating power need to be revised or replaced. The analysis results show that 8 items (32%) fall into the moderate category, 7 items (28%) fall into the difficult category, 3 items (12%) fall into the easy category, 6 items (24%) fall into the very easy category, and 1 item (4%) falls into the very difficult category. According to Purwanto and Angraini [11], the ideal distribution of difficulty levels is 25% easy items, 50% moderate items, and 25% difficult items.

Thus, the distribution of difficulty levels on multiple-choice items is still less than ideal, as the percentage of items with moderate difficulty level is still less than 50%. Based on the analysis results, 10 items (40%) have a very significant correlation with the total score, 10 items (40%) have a non-significant correlation with the total score, and 5 items (20%) cannot have their correlation calculated (NAN). Items with a very significant correlation show that these items can measure the same ability as the test as a whole. Conversely, items with a non-significant correlation show that these items cannot measure the same ability as the test as a whole.

According to Febrilia [12], items with a non-significant correlation need to be revised or replaced. The analysis results show that from 75 distractors (3 distractors on 25 items), 19 distractors (25.33%) function very well, 22 distractors (29.33%) function well, 3 distractors (4%) function fairly, 16 distractors (21.33%) function poorly, and 15 distractors (20%) function very poorly. Distractors that function well can mislead test takers who do not know the correct answer. Conversely, distractors that do not function well cannot mislead test takers who do not know the correct answer. Sutiasih and Saputri [13] suggest that distractors that do not function well need to be revised or replaced.

3.1.4. Quality of Essay Questions

The analysis results show that the reliability coefficient of the essay test is 0.93, which falls into the very high category. This indicates that the essay test has very high consistency in measuring students' abilities. Widodo and Kusnanik [14] state that high reliability on essay tests indicates that the test has good quality. Based on the analysis results, 6 items (60%) have good discriminating power, 2 items (20%) have fair discriminating power, and 2 items (20%) have poor discriminating power. Items with good and fair discriminating power show that these items can differentiate between high-ability test takers and low-ability test takers.

Conversely, items with poor discriminating power, namely items number 4 and 9, show that these items cannot differentiate between high-ability test takers and low-ability test takers. Basuki and Hariyanto [15] suggest that items with poor discriminating power need to be revised, especially in terms of construction and content of the material being tested. The analysis results show that all essay items (100%) fall into the moderate category. This indicates that the essay questions have an ideal difficulty level. According to Arifin [16], items with moderate difficulty levels can maximally describe students' abilities, as these items are neither too easy nor too difficult.

Nevertheless, for selection or ranking purposes, it is also necessary to include some items with varied difficulty levels to be able to differentiate students' abilities. Based on the analysis results, 8 items (80%) have a very significant correlation with the total score and 2 items (20%) cannot have their correlation calculated (NAN). Items with a very significant correlation show that these items can measure the same ability as the test as a whole. Meanwhile, items that cannot have their correlation calculated (items number 4 and 9) require further study to determine whether these items need to be retained or replaced.

4. CONCLUSIONS

Based on the ANATES analysis results on the evaluation instrument for 6th Grade PJOK subjects consisting of 25 multiple-choice items and 10 essay items, it can be concluded that the reliability of the multiple-choice questions has a reliability coefficient of 0.72, indicating a fairly good level of reliability. Meanwhile, the essay questions have a higher reliability coefficient of 0.93, indicating a very good level of reliability. Overall, essay questions have higher consistency than multiple-choice questions in measuring students' abilities.

Regarding difficulty levels in multiple-choice questions, there is a fairly diverse variation of difficulty levels: 7 very easy items (28%), 3 easy items (12%), 7 moderate items (28%), 7 difficult items (28%), and 1 very difficult item (4%). Meanwhile, for essay questions, all items (100%) are in the moderate category with a range of 33.33% - 66.67%. The distribution of difficulty levels for multiple-choice questions is less proportional with dominance of very easy and difficult questions, while essay questions have a more balanced distribution. Regarding discriminating power in multiple-choice questions, 11 items (44%) have good discriminating power (with a value of 100%), while 14 items (56%) have a discriminating power of 0%.

Meanwhile, for essay questions, 4 items (40%) have good discriminating power (66.67%), 4 items (40%) have fair discriminating power (33.33%), and 2 items (20%) have low discriminating power (0%). In general, many items still need to have their discriminating power improved to be able to differentiate between high-ability and low-ability students. Regarding correlation of item scores with total scores in multiple-choice questions, 10 items (40%) have a "Very Significant" correlation, 5 items (20%) are not significant, 1 item (4%) has a negative correlation, and 9 items (36%) cannot have their correlation calculated (NAN).

Meanwhile, for essay questions, 8 items (80%) have a "Very Significant" correlation and 2 items (20%) cannot have their correlation calculated (NAN). Essay questions show better item validity than multiple-choice questions. Regarding distractor quality (specifically for multiple-choice questions), there is a variation in distractor quality from very good to very poor. Several items have distractors that do not function well (too easily recognized as incorrect answers). Distractors in items categorized as poor and very poor need improvement to increase test effectiveness. Essay question instruments have better psychometric quality than multiple-choice questions in terms of reliability, distribution of difficulty levels, and item validity.

The very limited sample size (only 5 students) may affect the accuracy of the analysis, especially in calculating correlation and discriminating power. Several multiple-choice items need revision, especially those with low discriminating power, negative correlation with total scores, and poor distractor quality. For essay questions, items number 4 and 9 need special attention because they have a discriminating power of 0% and a NAN correlation.

REFERENCES

Mardapi, D., & Kartowagiran, B. (2021). *Pengukuran, penilaian, dan evaluasi pendidikan* (Edisi 3). UNY Press.

Azwar, S. (2021). *Reliabilitas dan Validitas* (Edisi 5). Pustaka Pelajar.

Blegur, J., & Lumba, A. (2022). Evaluasi pendidikan jasmani: Analisis kualitas instrumen tes dan pengukuran. *Jurnal Pendidikan Jasmani dan Olahraga*, 5(1), 25-36. https://doi.org/10.17509/jpjo.v5i1.31569

Dewi, P. C., & Mukminan, M. (2023). Analisis kualitas butir soal ujian sekolah mata pelajaran PJOK dengan pendekatan ANATES. *Jurnal Pendidikan dan Keolahragaan*, 8(1), 11-25.

Nuryadi, N., & Yanda, A. (2023). Software ANATES dan aplikasinya dalam analisis kualitas butir soal: Studi kasus pada mata pelajaran PJOK. *Indonesian Journal of Educational Assessment*, 6(1), 28-41.

Gunawan, I. (2020). *Metode penelitian: Kuantitatif, kualitatif, dan evaluasi* (Edisi 2). Universitas Negeri Malang.

Supriyadi, E. (2021). *Analisis Butir Soal dengan Software ANATES Versi 4 dan ITEMAN*. UNY Press.

Jannah, R. (2022). Analisis butir soal dengan bantuan aplikasi ANATES. Jurnal Ilmiah Pendidikan, 6(2), 124-135. https://doi.org/10.26858/jip.v6i2.28901

Sukiman. (2021). Pengembangan sistem evaluasi pendidikan. Insan Madani.

Arifin, Z., & Retnawati, H. (2022). Analisis butir soal: Konstruksi dan aplikasi dalam penilaian. Jurnal Penelitian dan Evaluasi Pendidikan, 26(1), 1-12. https://doi.org/10.21831/pep.v26i1.41270

Purwanto, A., & Angraini, R. (2020). Penggunaan software ANATES untuk validasi instrumen tes. Jurnal Ilmiah Pendidikan dan Pembelajaran, 4(2), 332-341. https://doi.org/10.23887/jipp.v4i2.27417

Febrilia, B. R. A. (2021). Analisis butir soal dalam evaluasi pendidikan: Konsep dasar penerapan. Jurnal Teknologi Pendidikan, 23(1), 67-78. https://doi.org/10.21009/jtp.v23i1.22440

Sutiasih, E., & Saputri, D. Y. (2022). Analisis butir soal ujian semester dengan program ANATES. Jurnal Pendidikan dan Kebudayaan, 7(1), 56-69. https://doi.org/10.24832/jpnk.v7i1.2704

Widodo, A., & Kusnanik, N. W. (2023). Evaluasi pembelajaran pendidikan jasmani di sekolah dasar. Jurnal Keolahragaan, 11(1), 87-98. https://doi.org/10.21831/jk.v11i1.51327

Basuki, I., & Hariyanto. (2020). Asesmen pembelajaran. PT Remaja Rosdakarya.

Arifin, Z. (2021). Evaluasi pembelajaran: Prinsip, teknik, dan prosedur. PT Remaja Rosdakarya.