IMEIJ
Indo-MathEdu Intellectuals Journal

p-ISSN: 2808 – 5604
e-ISSN: 2808 – 5078
Volume. 6, No. 4, 2025

ANALISIS BUTIR SOAL PILIHAN GANDA DENGAN PENDEKATAN CLASSICAL TEST THEORY PADA UJIAN AKHIR SEMESTER MATA PELAJARAN MATEMATIKA

Abdul Rahim[1], Yuliana Olga Siba Sabon[2], Nurhayati[3], Bhujangga Ayu Priyudahari[4], Rukmana Fachrul Islam[5], Haerul Amri[6]
[1, 2, 3, 4, 5, 6]Universitas Musamus, Jl. Kamizaun, Merauke, Papua, Indonesia
Email: abdul.rahim@unmus.ac.id

Article History

Received: 06-06-2025

Revision: 19-06-2025

Accepted: 23-06-2025

Published: 27-06-2025

Abstract. The purpose of this research is to analyze question items with a classical test theory approach. The type of research is descriptive quantitative. The object of this research is multiple choice items used in the end of semester examination of mathematics subjects in junior high school. The research subjects were students who took the test. The number of respondents was 50 people. The research instrument was 30 multiple choice questions. The data were analyzed using the Classical Test Theory approach with the help of RStudio analysis software. The results of the analysis of the level of difficulty of 30 items, it is known that as many as 28 items (93.33) are difficult categories. There are 2 items (6.67%) in the medium category. There is no single question that is classified as easy. Differentiability as many as 7 items (23.3%) in the category of acceptable with improvement. A total of 12 items (40%) in the corrected category. Meanwhile, 11 items (36.7%) were in the discarded category. All options A to E on each item have been selected, so the distractor function works well. The reliability of each item is in the range between 0.70 to 0.71 in the high reliability category. Overall, the reliability of the test instrument is 0.713, which is also in the high category.

Keywords: Item Analysis, Multiple Choice, Classical Test Theory, Final Semester Examination, Mathematics

Abstrak. Tujuan penelitian ini adalah menganalisis butir soal dengan pendekatan classical test theory. Jenis penelitian adalah deskriptif kuantitatif. Objek penelitian ini adalah butir soal pilihan ganda yang digunakan dalam ujian akhir semester mata pelajaran matematika pada sekolah menengah pertama. Subjek penelitian adalah peserta didik yang mengikuti tes. Jumlah responden sebanyak 50 orang. Instrumen penelitian adalah 30 butir soal pilihan ganda. Data dianalisis menggunakan pendekatan Classical Test Theory dengan bantuan software analisis RStudio. Hasil analisis tingkat kesukaran terhadap 30 butir soal, diketahui bahwa sebanyak 28 butir soal (93,33) kategori sulit. Terdapat 2 butir soal (6,67%) kategori sedang. Tidak terdapat satupun soal yang tergolong mudah. Daya beda sebayak 7 butir soal (23,3%) kategori diterima dengan perbaikan. Sebanyak 12 butir soal (40%) kategori diperbaiki. Sementara itu, 11 butir soal (36,7%) berada dalam kategori dibuang. Seluruh opsi A hingga E pada setiap butir soal telah dipilih maka fungsi distraktor bekerja dengan baik. Reliabilitas setiap butir berada pada rentang antara 0,70 hingga 0,71 dalam kategori reliabilitas tinggi. Secara keseluruhan, reliabilitas instrumen tes sebesar 0,713, yang juga termasuk dalam kategori tinggi.

Kata Kunci: Anaisis Butir Soal, Pilihan Ganda, Classical Test Theory, Ujian Akhir Semester, Matematika

How to Cite: Rahim, A., Sabon, Y. O. S., Nurhayati., Priyudahari, B. A., Islam, R. F., & Amri, H. (2025). Analisis Butir Soal Pilihan Ganda dengan Pendekatan *Classical Test Theory* pada Ujian Akhir Semester Mata Pelajaran Matematika. *Indo-MathEdu Intellectuals Journal*, 6 (4), 4763-4776. http://doi.org/10.54373/imeij.v6i4.3357

PENDAHULUAN

Evaluasi pembelajaran merupakan komponen integral dalam sistem pendidikan yang berfungsi untuk menilai pencapaian kompetensi peserta didik. Melalui evaluasi, dapat mengidentifikasi sejauh mana tujuan pembelajaran telah tercapai serta merancang tindak lanjut yang tepat (Gultom et al., 2024). Salah satu bentuk evaluasi yang paling umum digunakan adalah tes tertulis, yang dalam praktiknya menjadi acuan utama dalam pengambilan keputusan akademik terhadap peserta didik (Putri et al., 2024; Susanto, 2023). Namun, dalam praktik penyusunan soal ujian, masih banyak ditemui kendala terkait kualitas instrumen yang digunakan (Ayu et al., 2018; Zulrafli et al., 2023). Tidak sedikit soal yang disusun tanpa proses validasi dan analisis yang sistematis, sehingga berpotensi mengandung kelemahan seperti tingkat kesukaran yang tidak sesuai, daya pembeda yang rendah, serta distraktor yang tidak berfungsi optimal (Solichin, 2017; Yani et al., 2014). Kondisi ini dapat mengurangi keefektifan instrumen dalam mengukur kemampuan peserta didik secara objektif dan adil.

Soal yang tidak memenuhi kriteria kualitas secara psikometrik dapat menghasilkan data asesmen yang bias dan tidak mencerminkan kemampuan nyata peserta didik (Susongko & Muljani, 2024; Syadiah & Hamdu, 2020). Hal ini bukan hanya berdampak pada akurasi penilaian, tetapi juga dapat memengaruhi motivasi belajar siswa serta kepercayaan terhadap proses evaluasi itu sendiri. Oleh karena itu, perlu adanya upaya sistematis untuk menilai dan menyempurnakan soal melalui pendekatan ilmiah.

Salah satu pendekatan yang dapat digunakan untuk mengevaluasi kualitas butir soal adalah analisis psikometrik, khususnya dengan menelaah karakteristik soal secara kuantitatif (Susdelina et al., 2018). Melalui pendekatan ini, dapat diketahui informasi penting seperti tingkat kesukaran butir, daya pembeda, dan efektivitas distraktor (Akhmadi, 2021; Sari et al., 2021). Informasi tersebut berguna dalam merevisi, mengganti, atau mempertahankan soal sesuai dengan tujuan pengukuran yang diharapkan.

Classical Test Theory (CTT) merupakan pendekatan analisis soal yang banyak digunakan karena bersifat sederhana, mudah diterapkan, dan tidak membutuhkan perangkat lunak statistik yang kompleks (Erfan et al., 2020; Jafar & Ridwan, 2024). CTT menyediakan indikator-indikator utama yang dapat dijadikan dasar dalam menilai kualitas suatu soal, seperti nilai indeks kesukaran, indeks diskriminasi, dan analisis pengecoh. Oleh karena itu, CTT sangat relevan digunakan dalam konteks pendidikan, khususnya bagi guru dan dosen di berbagai jenjang.

Beberapa penelitian sebelumnya telah banyak membahas analisis butir soal berdasarkan pendekatan *Classical Test Theory*. Penelitian oleh (Fiska et al., 2021) yang menggunakan software Anates untuk menganalisis butir soal ulangan harian mata pelajaran IPA, serta penelitian oleh (Setyawarno, 2017) yang menerapkan aplikasi Iteman dalam menganalisis soal pilihan ganda. Selain itu, penelitian oleh (Purwati et al., 2021) menggunakan perangkat lunak QUEST untuk menganalisis karakteristik butir soal Ujian Nasional Matematika tingkat SMP/MTs. Berdasarkan peneltian sebelumnya kebaruan penelitian ini menggunakan software analisis R-Studio untuk menganalisis kualitas butir soal dengan pendekatan classical test theory. Urgensi dari penelitian ini terletak pada pentingnya membangun budaya evaluasi soal yang berbasis data dan analisis ilmiah. Dengan mendorong penggunaan analisis CTT dalam evaluasi butir soal, diharapkan diperoleh instrumen yang valid, reliabel, dan adil. Penelitian ini juga diharapkan dapat menjadi rujukan praktis dalam pengembangan soal yang bermutu dalam rangka peningkatan kualitas pembelajaran secara menyeluruh.

METODE

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan deskriptif. Tujuan dari penelitian ini adalah untuk menganalisis kualitas butir soal berdasarkan karakteristik statistiknya, yaitu tingkat kesukaran, daya pembeda, dan efektivitas distraktor, sesuai dengan prinsip-prinsip Classical Test Theory (CTT). Objek penelitian ini adalah butir soal pilihan ganda yang digunakan dalam ujian akhir semester mata pelajaran matematika pada sekolah menengah pertama. Subjek penelitian adalah peserta didik yang mengikuti tes tersebut, yang hasil tesnya dianalisis secara statistik. Jumlah responden sebanyak 50 orang.

Instrumen utama dalam penelitian ini adalah dokumen soal pilihan ganda beserta lembar jawaban peserta didik. Instrumen bantu berupa pedoman analisis butir soal dan perangkat lunak statistik RStudio untuk mengolah data hasil tes. Data dikumpulkan melalui dokumentasi, yaitu dengan mengumpulkan naskah soal dan lembar jawaban siswa. Data yang dikumpulkan mencakup kunci jawaban, skor setiap peserta didik, dan rekapitulasi jawaban per butir soal. Data dianalisis menggunakan pendekatan Classical Test Theory (CTT) dengan bantuan software analisis RSudio dengan menghitung tiga parameter utama (tingkat kesukran, daya beda, distraktor) dan reliabilitas.

Tingkat Kesukaran (TK)

Tingkat Kesukaran (Difficulty Index) untuk mengukur proporsi peserta yang menjawab benar setiap butir soal. Adapun rumus untuk menghitung Tingkat Kesuakran (TK) dan Kriteria

Tabel 1. Rumus menghitung dan kriteria Tingkat Kesukaran (TK)

Rumus Menghitung Tingkat Kesukaran (TK)	Kriteria Tingkat Kesukaran (TK)
$$\boldsymbol{TK = \frac{BA + BB}{N}}$$ TK : Tingkat Kesulitan BA : jumlah jawaban benar kelas atas BB : jumlah jawaban benar kelas bawah N : jumlah keselurahan testee	▪ **> 0,7** = Mudah ▪ **0,3 – 0,7** = Sedang ▪ **< 0,3** = Sulit

Daya Beda (DB)

Daya Pembeda (Discrimination Index) untuk mengukur kemampuan soal dalam membedakan antara peserta dengan kemampuan tinggi dan rendah. Adapun rumus untuk Daya Beda (DB) dan Kriteria Daya Beda dapat dilihat pada Tabel 2.

Tabel 2. Rumus Menghitung dan Kriteria Daya Beda (DB)

Rumus Menghitung Daya Beda (DB)	Kriteria Daya Beda (DB)
$$DB = \frac{BA - BB}{\frac{1}{2}N}$$ DB : Daya Beda BA : jumlah jawaban benar kelas atas BB : jumlah jawaban benar kelas bawah N : jumlah keselurahan testee	▪ 0,40 – 1,00 = Diterima Baik ▪ 0,30 – 0,39 = Diterima (Perlu Perbaiki) ▪ 0,20 – 0,29 = Diperbaiki ▪ 0,00 – 0,19 = Dibuang

Distraktor

Efektivitas Distraktor untuk menilai sejauh mana pengecoh (opsi jawaban yang salah) berfungsi menarik perhatian peserta yang kurang memahami materi. Distraktor berfungdi dengan baik jika ada subjek yang memilih (Nitko, 1996).

Reliabilitas

Reliabilitas butir soal merujuk pada konsistensi atau keajegan hasil jawaban terhadap suatu butir soal dalam suatu instrumen tes. Dengan kata lain, reliabilitas butir soal menunjukkan sejauh mana sebuah item (soal) dapat memberikan hasil yang stabil dan dapat dipercaya jika digunakan dalam pengukuran yang berulang. Adapun rumus menghitung Reliabilita Alpha Cronbach dan Kriteria tingkat Reliabilitas tes pada Tabel 3.

Tabel 3. Rumus Menghitung dan Kriteria Reliabilitas

Rumus Menghitung Reliabilitas	Kriteria Reliabilitas

$$r = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum \sigma^2 b}{\sigma^2 t}\right]$$

r : Koofisien Reliabilitas Alpha
k : jumlah item pertanyaan
$\sum \sigma^2 b$: jumlah varian butir
$\sigma^2 t$: Varian total

- 0,80 – 1,00 = Sangat Tinggi
- 0,60 – 0,79 = Tinggi
- 0,40 – 0,59 = Cukup
- 0,20 – 0,39 = Rendah
- 0,00 – 0,19 = Sangat Rendah

HASIL DAN DISKUSI

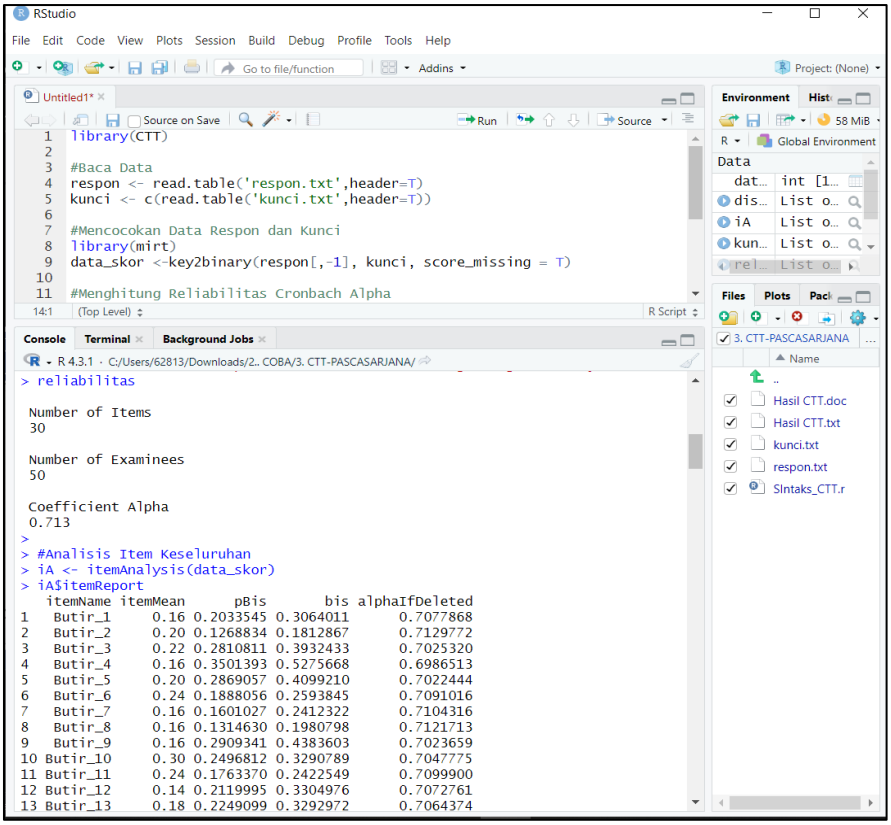
Skrip Analisis *Classical Test Theory* (CTT) dengan RStudio

Untuk menganalisis kualitas butir soal dan reliabilitas secara kuantitatif, digunakan pendekatan *Classical Test Theory* (CTT) melalui software analisis RStudio. Dalam proses ini, digunakan skrip analisis dengan bantuan pustaka *CTT* dan *mirt* yang memungkinkan perhitungan analisis tingkat kesukaran butir, daya beda butir, distraktor, serta reliabilitas Cronbach's Alpha. Skrip dan hasil analisis *Classical Test Theory* (CTT) dengan RStudio dapat dilihat pada Gambar 1 dan Gambar 2.

SIntaks_CTT.r

```
1
2   library(CTT)
3
4   #Baca Data
5   respon <- read.table('respon.txt',header=T)
6   kunci <- c(read.table('kunci.txt',header=T))
7
8   #Mencocokan Data Respon dan Kunci
9   library(mirt)
10  data_skor <-key2binary(respon[,-1], kunci, score_missing = T)
11
12  #Menghitung Reliabilitas Cronbach Alpha
13  reliabilitas <- reliability(data_skor)
14  reliabilitas
15
16  #Analisis Item Keseluruhan
17  iA <- itemAnalysis(data_skor)
18  iA$itemReport
19
20  #Analisis Item Satu per Satu
21  distraktor <- distractorAnalysis(respon[,-1], kunci)
22  distraktor
23
24  #Simpan Hasil Analisis
25  options(width=300)
26  sink('Hasil CTT.txt', split = T)
27  cat('\nReliabilitas\n')
28  reliabilitas
29  cat('\nItem Analisis\n')
30  iA$itemReport
31  cat('\nDistraktor Analisis\n')
32  distraktor
33  sink()
34
```

Gambar 1. Skrip analisis *Classical Test Theory* (CTT)



Gambar 2. Tampilan Analisis *Classical Test Theory* (CTT) dengan RStudio

Tingkat Kesukaran

Tingkat kesukaran butir soal adalah ukuran yang menunjukkan sejauh mana suatu soal dapat dijawab dengan benar oleh peserta tes. Indeks ini mencerminkan proporsi peserta yang menjawab benar terhadap suatu butir, sehingga dapat digunakan untuk menilai apakah soal tersebut tergolong mudah, sedang, atau sulit. Secara umum, tingkat kesukaran berperan penting dalam menentukan kualitas dan keseimbangan keseluruhan suatu instrumen tes.

Tabel 4. Hasil analisis tingkat kesukaran butir soal

Butir Soal	Tingkat Kesukaran (TK)	Kriteria Tingkat Kesukaran (TK)
1	0,16	Sulit
2	0,2	Sulit
3	0,22	Sulit
4	0,16	Sulit
5	0,2	Sulit
6	0,24	Sulit
7	0,16	Sulit
8	0,16	Sulit
9	0,16	Sulit
10	0,3	Sedang
11	0,24	Sulit
12	0,14	Sulit

13	0,18	Sulit
14	0,28	Sulit
15	0,2	Sulit
16	0,28	Sulit
17	0,14	Sulit
18	0,26	Sulit
19	0,22	Sulit
20	0,3	Sedang
21	0,18	Sulit
22	0,2	Sulit
23	0,16	Sulit
24	0,16	Sulit
25	0,24	Sulit
26	0,24	Sulit
27	0,2	Sulit
28	0,2	Sulit
29	0,16	Sulit
30	0,18	Sulit

Berdasarkan Tabel 4 hasil analisis tingkat kesukaran terhadap 30 butir soal, diketahui bahwa mayoritas soal tergolong dalam kategori sukar. Sebanyak 28 butir soal atau sebesar 93,33% memiliki indeks kesukaran di bawah 0,30, yang menandakan bahwa soal-soal tersebut terlalu sulit bagi sebagian besar peserta tes. Hanya terdapat 2 butir soal (6,67%) yang masuk dalam kategori sedang, yaitu Butir 10 dan Butir 20, dengan indeks kesukaran masing-masing sebesar 0,30. Tidak terdapat satupun soal yang tergolong mudah. Kondisi ini menunjukkan bahwa proporsi tingkat kesukaran dalam instrumen soal belum seimbang, yang dapat berdampak negatif terhadap reliabilitas dan daya pembeda tes. Soal yang terlalu sulit secara dominan dapat menyebabkan peserta kesulitan dalam menjawab dengan benar, sehingga hasil tes tidak merefleksikan kemampuan peserta secara akurat.

Daya Beda

Daya beda butir soal adalah suatu indikator dalam analisis tes yang menunjukkan kemampuan suatu butir soal dalam membedakan peserta didik yang memiliki kemampuan tinggi dan rendah. Butir yang memiliki daya beda baik akan dijawab benar oleh peserta dengan kemampuan tinggi dan dijawab salah oleh peserta dengan kemampuan rendah. Dalam konteks evaluasi pembelajaran, daya beda yang tinggi mencerminkan bahwa suatu soal bersifat diskriminatif dan mampu berkontribusi secara signifikan terhadap kualitas instrumen tes secara keseluruhan. Hasil analisis Daya Beda dapat dilihat pada Tabel 5.

Tabel 5. Hasil analisis daya beda butir soal

Butir Soal	Daya Beda (DB)	Kriteria Daya Beda (DB)
1	0,20	Diperbaiki
2	0,13	Dibuang
3	0,28	Diperbaiki
4	0,35	Diterima (Perlu Diperbaiki)
5	0,29	Diperbaiki
6	0,19	Dibuang
7	0,16	Dibuang
8	0,13	Dibuang
9	0,29	Diterima (Perlu Diperbaiki)
10	0,25	Diperbaiki
11	0,18	Dibuang
12	0,21	Diperbaiki
13	0,22	Diperbaiki
14	0,19	Dibuang
15	0,30	Diterima (Perlu Diperbaiki)
16	0,12	Dibuang
17	0,23	Diperbaiki
18	0,16	Dibuang
19	0,37	Diterima (Perlu Diperbaiki)
20	0,21	Diperbaiki
21	0,17	Dibuang
22	0,25	Diperbaiki
23	0,38	Diterima (Perlu Diperbaiki)
24	0,31	Diterima (Perlu Diperbaiki)
25	0,26	Diperbaiki
26	0,19	Dibuang
27	0,14	Dibuang
28	0,27	Diperbaiki
29	0,35	Diterima (Perlu Diperbaiki)
30	0,27	Diperbaiki

Berdasarkan Tabel 5 hasil analisis terhadap 30 butir soal, diketahui bahwa kualitas soal berdasarkan daya beda menunjukkan variasi yang cukup signifikan. Sebanyak 7 butir soal (23,3%) berada dalam kategori diterima dengan perbaikan, artinya soal-soal ini memiliki kemampuan yang baik dalam membedakan peserta didik berkemampuan tinggi dan rendah, namun tetap perlu dilakukan revisi ringan agar lebih optimal. Sebanyak 12 butir soal (40%) termasuk dalam kategori diperbaiki, yaitu memiliki daya beda sedang dan masih dapat digunakan setelah diperbaiki baik dari segi redaksional maupun substansi. Sementara itu, 11 butir soal (36,7%) berada dalam kategori dibuang karena nilai daya bedanya kurang dari 0,20, menunjukkan bahwa soal-soal ini tidak mampu membedakan peserta dengan kemampuan tinggi dan rendah secara efektif.

Distraktor

Fungsi distraktor adalah peran pilihan jawaban yang tidak benar dalam butir soal pilihan ganda yang dirancang untuk menarik peserta didik yang belum memahami materi dengan baik. Distraktor yang baik harus cukup menarik sehingga dipilih oleh peserta dengan kemampuan rendah, namun dihindari oleh peserta dengan kemampuan tinggi. Secara ilmiah, fungsi distraktor yang efektif berkontribusi terhadap validitas butir soal karena menunjukkan bahwa setiap alternatif jawaban memiliki fungsi diagnostik dalam mengungkap tingkat pemahaman peserta.

Tabel 6. Fungsi distraktor butir soal

Butir Soal	Pilihan Jawaban				
	A	B	C	D	E
1	24%	18%	26%	16%	16%
2	20%	20%	20%	12%	28%
3	22%	14%	12%	22%	30%
4	20%	28%	20%	16%	16%
5	18%	18%	18%	20%	26%
6	16%	18%	20%	22%	24%
7	10%	24%	16%	34%	16%
8	24%	16%	16%	28%	16%
9	16%	24%	14%	16%	30%
10	26%	10%	12%	22%	30%
11	16%	24%	18%	24%	18%
12	20%	24%	20%	14%	22%
13	22%	18%	22%	22%	16%
14	28%	14%	24%	14%	20%
15	20%	20%	28%	20%	12%
16	6%	30%	28%	24%	12%
17	26%	20%	14%	22%	18%
18	18%	18%	12%	26%	26%
19	22%	18%	26%	18%	16%
20	22%	10%	16%	22%	30%
21	28%	14%	18%	24%	16%
22	16%	14%	20%	30%	20%
23	16%	18%	22%	18%	26%
24	16%	24%	24%	22%	14%
25	18%	22%	12%	24%	24%
26	20%	24%	16%	16%	24%
27	20%	18%	20%	24%	18%
28	22%	22%	10%	20%	26%
29	24%	16%	30%	14%	16%
30	10%	12%	22%	38%	18%

Berdasarkan Tabel 6 distribusi pilihan jawaban, seluruh opsi A hingga E pada setiap butir soal telah dipilih oleh peserta, yang menunjukkan bahwa tidak ada distraktor yang diabaikan. Dengan demikian, dapat disimpulkan bahwa fungsi distraktor bekerja dengan baik karena semua pilihan jawaban cukup menarik dan mampu mengecoh peserta yang belum memahami materi secara utuh.

Reliabilitas

Reliabilitas adalah tingkat konsistensi atau keterandalan suatu instrumen dalam mengukur apa yang seharusnya diukur secara berulang dalam kondisi yang relatif sama. Dalam konteks evaluasi pembelajaran, reliabilitas menunjukkan sejauh mana hasil tes dapat dipercaya dan bebas dari kesalahan acak. Instrumen yang memiliki reliabilitas tinggi akan menghasilkan skor yang stabil dan konsisten, serta mencerminkan kualitas pengukuran yang baik dalam menilai kemampuan peserta didik.

Tabel 7. Reliabilitas

Butir Soal	Reliabilitas	Kriteria
1	0,71	Tinggi
2	0,71	Tinggi
3	0,70	Tinggi
4	0,70	Tinggi
5	0,70	Tinggi
6	0,71	Tinggi
7	0,71	Tinggi
8	0,71	Tinggi
9	0,70	Tinggi
10	0,70	Tinggi
11	0,71	Tinggi
12	0,71	Tinggi
13	0,71	Tinggi
14	0,71	Tinggi
15	0,70	Tinggi
16	0,71	Tinggi
17	0,71	Tinggi
18	0,71	Tinggi
19	0,70	Tinggi
20	0,71	Tinggi
21	0,71	Tinggi
22	0,70	Tinggi
23	0,70	Tinggi
24	0,70	Tinggi
25	0,70	Tinggi
26	0,71	Tinggi
27	0,71	Tinggi

28	0,70	Tinggi
29	0,70	Tinggi
30	0,70	Tinggi

Berdasarkan hasil analisis reliabilitas terhadap 30 butir soal, diperoleh nilai reliabilitas setiap butir berada pada rentang antara 0,70 hingga 0,71. Seluruh butir soal termasuk dalam kategori reliabilitas tinggi, yang menunjukkan bahwa setiap butir memiliki konsistensi yang baik dalam mengukur kemampuan peserta didik. Hal ini menandakan bahwa soal-soal tersebut relatif stabil dan dapat memberikan hasil yang andal jika digunakan kembali dalam kondisi serupa. Tidak terdapat butir dengan reliabilitas rendah atau sedang, sehingga secara umum kualitas soal tergolong sangat baik dari sisi keandalan. Secara keseluruhan, reliabilitas instrumen tes sebesar 0,713, yang juga termasuk dalam kategori tinggi. Reliabilitas instrumen tes mencerminkan keseluruhan tes memiliki tingkat konsistensi internal yang kuat, di mana antar butir soal saling mendukung dalam mengukur konstruk yang sama.

Integrasi Hasil Analisis

Integrasi hasil analisis tingkat kesukaran, daya beda, fungsi distraktor, dan reliabilitas memberikan gambaran komprehensif terhadap kualitas butir soal dalam suatu instrumen evaluasi. Tingkat kesukaran menunjukkan sejauh mana soal dapat dijawab dengan benar oleh peserta, daya beda mengukur kemampuan soal dalam membedakan peserta berkemampuan tinggi dan rendah, sedangkan fungsi distraktor menilai efektivitas pilihan jawaban yang salah dalam menarik peserta yang belum memahami materi. Ketika keempat aspek ini dianalisis secara terpadu, diperoleh informasi yang mendalam mengenai validitas dan keandalan instrumen, sehingga dapat digunakan sebagai dasar dalam merevisi, memperbaiki, atau mempertahankan butir soal untuk meningkatkan kualitas pengukuran hasil belajar.

Tabel 8. Integrasi hasil analisis

Butir	Tingkat Kesukaran (TK)	Daya Beda (BD)	Pilihan Jawaban					Saran Terhadap Butir	Reliabilitas
			A	B	C	D	E		
1	0,16	0,20	24 %	18 %	26 %	16 %	16 %	Diperbaiki	0,713
2	0,2	0,13	20 %	20 %	20 %	12 %	28 %	Dibuang	
3	0,22	0,28	22 %	14 %	12 %	22 %	30 %	Diperbaiki	
4	0,16	0,35	20 %	28 %	20 %	16 %	16 %	Diterima (Perlu Diperbaiki)	
5	0,2	0,29	18 %	18 %	18 %	20 %	26 %	Diperbaiki	
6	0,24	0,19	16 %	18 %	20 %	22 %	24 %	Dibuang	

7	0,16	0,16	10 %	24 %	16 %	34 %	16 %	Dibuang
8	0,16	0,13	24 %	16 %	16 %	28 %	16 %	Dibuang
9	0,16	0,29	16 %	24 %	14 %	16 %	30 %	Diterima (Perlu Diperbaiki)
10	0,3	0,25	26 %	10 %	12 %	22 %	30 %	Diperbaiki
11	0,24	0,18	16 %	24 %	18 %	24 %	18 %	Dibuang
12	0,14	0,21	20 %	24 %	20 %	14 %	22 %	Diperbaiki
13	0,18	0,22	22 %	18 %	22 %	22 %	16 %	Diperbaiki
14	0,28	0,19	28 %	14 %	24 %	14 %	20 %	Dibuang
15	0,2	0,30	20 %	20 %	28 %	20 %	12 %	Diterima (Perlu Diperbaiki)
16	0,28	0,12	6 %	30 %	28 %	24 %	12 %	Dibuang
17	0,14	0,23	26 %	20 %	14 %	22 %	18 %	Diperbaiki
18	0,26	0,16	18 %	18 %	12 %	26 %	26 %	Dibuang
19	0,22	0,37	22 %	18 %	26 %	18 %	16 %	Diterima (Perlu Diperbaiki)
20	0,3	0,21	22 %	10 %	16 %	22 %	30 %	Diperbaiki
21	0,18	0,17	28 %	14 %	18 %	24 %	16 %	Dibuang
22	0,2	0,25	16 %	14 %	20 %	30 %	20 %	Diperbaiki
23	0,16	0,38	16 %	18 %	22 %	18 %	26 %	Diterima (Perlu Diperbaiki)
24	0,16	0,31	16 %	24 %	24 %	22 %	14 %	Diterima (Perlu Diperbaiki)
25	0,24	0,26	18 %	22 %	12 %	24 %	24 %	Diperbaiki
26	0,24	0,19	20 %	24 %	16 %	16 %	24 %	Dibuang
27	0,2	0,14	20 %	18 %	20 %	24 %	18 %	Dibuang
28	0,2	0,27	22 %	22 %	10 %	20 %	26 %	Diperbaiki
29	0,16	0,35	24 %	16 %	30 %	14 %	16 %	Diterima (Perlu Diperbaiki)
30	0,18	0,27	10 %	12 %	22 %	38 %	18 %	Diperbaiki

Berdasarkan hasil analisis butir soal, terdapat 11 butir soal yang direkomendasikan untuk dibuang, yaitu Butir 2, 6, 7, 8, 11, 14, 16, 18, 21, 26, dan 27. Kesebelas butir tersebut memiliki nilai daya beda yang sangat rendah, berkisar antara 0,12 hingga 0,19, serta sebagian besar

menunjukkan tingkat kesukaran yang terlalu tinggi (soal terlalu sulit), sehingga tidak mampu membedakan antara peserta yang memiliki kemampuan tinggi dan rendah. Sementara itu, terdapat 19 butir soal yang masih dapat digunakan setelah melalui proses perbaikan dan penyempurnaan. Dari jumlah tersebut, 7 butir soal dinyatakan layak digunakan dengan catatan perlu perbaikan (Butir 4, 9, 15, 19, 23, 24, dan 29) karena memiliki daya beda yang cukup baik. Sedangkan 12 butir lainnya disarankan untuk diperbaiki (Butir 1, 3, 5, 10, 12, 13, 17, 20, 22, 25, 28, dan 30) agar dapat meningkatkan kualitas pengukuran, khususnya pada aspek daya beda. Dengan perbaikan yang tepat, butir-butir ini dapat menjadi bagian dari instrumen yang valid dan reliabel untuk mengukur capaian pembelajaran peserta. Secara keseluruhan, reliabilitas instrumen tes sebesar 0,713, yang termasuk dalam kategori tinggi.

KESIMPULAN

Berdasarkan hasil dan pembahasan dapat disimpulkan tingkat kesukaran terhadap 30 butir soal, diketahui bahwa mayoritas soal tergolong dalam kategori sukar. Sebanyak 28 butir soal atau sebesar 93,33% memiliki indeks kesukaran di bawah 0,30, yang menandakan bahwa soal-soal tersebut terlalu sulit bagi sebagian besar peserta tes. Hanya terdapat 2 butir soal (6,67%) yang masuk dalam kategori sedang, yaitu Butir 10 dan Butir 20, dengan indeks kesukaran masing-masing sebesar 0,30. Tidak terdapat satupun soal yang tergolong mudah.

Berdasarkan daya beda menunjukkan variasi yang cukup signifikan. Sebanyak 7 butir soal (23,3%) berada dalam kategori diterima dengan perbaikan. Sebanyak 12 butir soal (40%) termasuk dalam kategori diperbaiki. Sementara itu, 11 butir soal (36,7%) berada dalam kategori dibuang karena nilai daya bedanya kurang dari 0,20. Seluruh opsi A hingga E pada setiap butir soal telah dipilih oleh peserta, yang menunjukkan bahwa tidak ada distraktor yang diabaikan sehingga dapat disimpulkan bahwa fungsi distraktor bekerja dengan baik karena semua pilihan jawaban cukup menarik dan mampu mengecoh peserta yang belum memahami materi secara utuh.

Berdasarkan hasil analisis reliabilitas terhadap 30 butir soal, diperoleh nilai reliabilitas setiap butir berada pada rentang antara 0,70 hingga 0,71 berada dalam kategori reliabilitas tinggi. Secara keseluruhan, reliabilitas instrumen tes sebesar 0,713, yang juga termasuk dalam kategori tinggi.

REFERENSI

Akhmadi, M. N. (2021). Analisis Butir Soal Evaluasi Tema 1 Kelas 4 SDN Plumbungan Menggunakan Program Anates. *Ed-Humanistics*, *6*(1), 799–806.

Ayu, P. E. S., Marhaeni, A., & Adnyana, P. B. (2018). Pengembangan Instrumen Asesmen Keterampilan Belajar Dan Berinovasi Pada Mata Pelajaran Ipa Sd. *PENDASI: Jurnal Pendidikan Dasar Indonesia*, *2*(2), 90–100. https://doi.org/10.23887/jpdi.v2i2.2696

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis Kualitas Soal Kemampuan Membedakan Rangkaian Seri dan Paralel Melalui Teori Tes Klasik Dan Model Rasch. *Indonesian Journal of Educational Research and Review*, *3*(1), 11–19. https://doi.org/10.23887/ijerr.v3i1.24080

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

Gultom, Y. M., Syahputra, F., & Syahrial, S. (2024). Pengaruh Evaluasi Pembelajaran terhadap Kualitas Pembelajaran Guru di Sekolah Dasar. *Jurnal Pendidikan Guru Sekolah Dasar*, *1*(3), 1–8. https://doi.org/10.47134/pgsd.v1i3.543

Jafar, L., & Ridwan, A. (2024). Penggunaan Teori Tes Klasik Untuk Analisis Butir Soal Pada Asesmen. *IJSH: Indonesian Journal of Social and Humanities*, *02*(03), 1–10.

Purwati, H., Retnawati, H., Jailani, J., & Retnowati, T. H. (2021). Analisis Karakteristik Butir Soal Ujian Nasional Matematika SMP/MTs Berdasarkan Pendekatan Teori Tes Klasik. *Jurnal Sains Dan Edukasi Sains*, *4*(2), 46–51. https://doi.org/10.24246/juses.v4i2p46-51

Putri, H., Susiani, D., Wandani, N. S., & Putri, F. A. (2024). Instrumen Penilaian Hasil Pembelajaran Kognitif Pada Tes Objektif. *Jurnal Pendidikan Dan Ilmu Sosial (Jupendis)*, *2*(4), 86–96. https://doi.org/10.54066/jupendis.v2i4.2159

Sari, D. K., Sarah, S., & Mursyidi, W. (2021). Kualitas butir soal penilaian akhir semester (pas) yang disusun guru madrasah 1. *JIPIS Volume*, *30*(2), 57–69.

Setyawarno, D. (2017). Penggunaan Aplikasi Software Iteman (Item and Test Analysis) untuk Analisis Butir Soal Pilihan Ganda Berdasarkan Teori Tes Klasik. *Jurnal Ilmu Fisika Dan Pembelajarannya (JIFP)*, *1*(1), 11–21. https://doi.org/10.19109/jifp.v1i1.866

Solichin, M. (2017). Analisis Daya Beda Soal Taraf Kesukaran,Butir Tes, Validitas Butir Tes, Interpretasi Hasil Tes Valliditas Ramalan dalam Evaluasi Pendidikan. *Jurnal Manajemen Dan Pendidikan Islam 2*, *2*(2), 192–213. https://doi.org/10.26594/dirasat.v2i2.879

Susanto, S. (2023). Pengembangan Alat Dan Teknik Evaluasi Tes DalamPendidikan. *Jurnal Tarbiyah Jamiat Kheir*, *1*(1), 51–60.

Susdelina, Perdana, S. A., & Febrian. (2018). Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik Dan Rasch Model. *Jurnal Kiprah*, *6*(1), 41–48. https://doi.org/10.31629/kiprah.v6i1.574

Susongko, P., & Muljani, S. (2024). Model Asesmen Kelulusan Fase B pada Mata Pelajaran Bahasa Indonesia. *Journal of Education Research*, *5*(4), 6383–6390.

Syadiah, A. N., & Hamdu, G. (2020). Analisis rasch untuk soal tes berpikir kritis pada pembelajaran STEM di sekolah dasar. *Premiere Educandum: Jurnal Pendidikan Dasar Dan Pembelajaran*, *10*(December), 138–148. https://doi.org/10.25273/pe.v10i2.6524

Yani, A., Asri, A. F., & Burhan, A. (2014). Distraktor Soal Ujian Semester Ganjil Mata Pelajaran Produktif Di Smk Negeri 1 Indralaya Utara. *Jurnal Pendidikan Teknik Mesin*, *1*(2), 98–115.

Zulrafli, Kamarudin, & Erawati, Y. (2023). Peningkatan Kompetensi Dan Kreativitas Guru Melalui Pelatihan Pembuatan Soal-Soal Berbasis Higher Order Thingking Skill (HOTS) Pada Kelompok Kerja Guru (KKG) Penjas. *Journal of Human And Education*, *3*(3), 241–248.