

Analisis Validitas dan Reliabilitas Butir Soal Matematika Materi Aritmetika Sosial di SMPN 1 Lembah Sorik Merapi

Mislaini^{1*}, Adelina², Feriansyah Siregar³, Rina Sari Lubis⁴, Hakmi Rais Fauzan⁵

^{1,2,3,4}Program Studi Tadris Matematika, STAIN Mandailing Natal, Indonesia

⁵Pendidikan Matematika, STKIP Al Hikmah Surabaya, Indonesia

Email: mislaininasution@gmail.com

Abstract

This study aims to analyze the validity, reliability, difficulty level, and discriminating power of multiple-choice test items in the Arithmetic topic for eighth-grade students at SMPN 1 Lembah Sorik Merapi in the 2024/2025 academic year. The research employed a quantitative descriptive method involving 20 eighth-grade students as participants. Data were obtained from students' achievement test sheets and analyzed using Microsoft Excel. The Pearson Product Moment correlation was used to test validity, while Cronbach's Alpha was applied to determine reliability. The results show that out of 10 test items, 7 items are valid and 3 items are invalid. The reliability coefficient (R_{11}) indicates a very high level of internal consistency, confirming the instrument's reliability. The difficulty index reveals that 30% of the items are easy, 50% moderate, and 20% difficult, while the discrimination index shows that 60% of items are good, 10% fair, and 30% poor. Overall, the test instrument demonstrates good quality and is suitable for use as an evaluation tool in mathematics learning, although some items require revision to achieve better construct alignment and measurement precision.

Keywords: Validity, Reliability, Difficulty Level, Discrimination Index, Learning Evaluation

Abstrak

Penelitian ini bertujuan untuk menganalisis validitas, reliabilitas, tingkat kesukaran, dan daya pembeda butir soal pilihan ganda pada materi Aritmetika kelas VIII di SMPN 1 Lembah Sorik Merapi tahun ajaran 2024/2025. Jenis penelitian yang digunakan adalah deskriptif kuantitatif, dengan subjek sebanyak 20 siswa kelas VIII. Data dikumpulkan melalui instrumen tes hasil belajar, yang kemudian dianalisis menggunakan Microsoft Excel. Analisis validitas dilakukan dengan rumus korelasi Product Moment Pearson, sedangkan reliabilitas dihitung dengan rumus Alpha Cronbach. Hasil penelitian menunjukkan bahwa dari 10 butir soal, terdapat 7 soal valid dan 3 soal tidak valid. Nilai reliabilitas (R_{11}) menunjukkan tingkat konsistensi yang sangat tinggi, sehingga instrumen dapat dikategorikan reliabel. Analisis tingkat kesukaran menunjukkan 30% soal mudah, 50% sedang, dan 20% sukar, sedangkan analisis daya pembeda menunjukkan 60% soal berkategori baik, 10% cukup, dan 30% jelek. Secara keseluruhan, instrumen tes dinilai memiliki kualitas yang baik dan layak digunakan sebagai alat evaluasi pembelajaran matematika, meskipun beberapa butir perlu direvisi agar mencapai kesesuaian konstruk yang lebih optimal.

Kata kunci: Validitas, Reliabilitas, Tingkat Kesukaran, Daya Pembeda, Evaluasi Pembelajaran.

How to Cite: Mislaini et. al. (2026). Analisis Kualitas Butir Soal Matematika Realistik Pada Mahasiswa Tadris Matematika STAIN Mandailing Natal. *Journal of Mathematics in Teaching and Learning*, 4 (2), 452-464.

PENDAHULUAN

Dalam institusi pendidikan, terutama di lembaga formal, yang paling penting adalah pembelajaran, evaluasi adalah bagian penting dari pembelajaran sebagai penilaian akhir semester yang merupakan salah satu jenis evaluasi yang paling sering dilakukan oleh institusi pendidikan (Kurniawan, 2019; Rahmi et al., 2024). Menurut Arikunto (2021), evaluasi adalah proses sistematis yang digunakan untuk mengevaluasi dan menentukan tingkat pencapaian siswa terhadap tujuan pembelajaran yang telah ditetapkan. Evaluasi erat kaitannya pula dengan penilaian dan pengukuran. Sedangkan menurut Brookhart & Nitko (2019) menegaskan bahwa evaluasi yang baik harus mampu memberikan informasi

Received December 26, 2025; Revised January 09, 2026; Accepted June 30, 2026

yang relevan bagi pengambilan keputusan pendidikan, baik untuk kebutuhan diagnosis, perbaikan, maupun penilaian hasil belajar siswa. Oleh karena itu, evaluasi menempati posisi penting dalam menjamin mutu pendidikan.

Evaluasi pembelajaran memiliki keterkaitan yang erat dengan dua konsep utama lainnya, yaitu pengukuran dan penilaian (Sudjana, 1995). Ketiga konsep ini memiliki hubungan yang bersifat hierarkis. Menurut Arikunto (2021), pengukuran menghasilkan data kuantitatif, penilaian menginterpretasikan data tersebut secara kualitatif, sedangkan evaluasi merupakan proses menyimpulkan hasil penilaian untuk mengambil keputusan pendidikan. Hal ini sejalan dengan pendapat Popham (2003) yang menyatakan bahwa evaluasi adalah “the process of making judgments based on measurement results”, yaitu proses membuat keputusan berdasarkan hasil pengukuran yang dilakukan secara objektif. Sedangkan menurut, Anderson (2005) pengukuran berfungsi sebagai dasar ilmiah dalam menentukan nilai yang digunakan dalam evaluasi pendidikan. Evaluasi didahului dengan penilaian, sedangkan penilaian didahului dengan pengukuran. Tes adalah cara yang digunakan secara tidak langsung untuk mengukur kemampuan seseorang melalui respons mereka terhadap rangsangan atau pertanyaan yang sudah mereka jawab (Mardapi, 2012).

Dalam praktik pendidikan di sekolah, salah satu bentuk evaluasi yang sering digunakan adalah tes sumatif. Tes ini dilakukan pada akhir pembelajaran atau akhir semester untuk mengukur hasil belajar secara menyeluruh (Arikunto, 2021). Tes sumatif bertujuan memberikan gambaran komprehensif tentang sejauh mana peserta didik memahami, menguasai, dan mampu menerapkan materi pelajaran yang telah diajarkan (Hayati et al., 2024; Zetra & Utami, 2025). Bentuk tes yang umum digunakan dalam tes sumatif adalah soal pilihan ganda, karena mampu mencakup banyak indikator kemampuan dengan waktu pengerjaan yang efisien dan hasil yang mudah diinterpretasikan (Sudjana, 1995). Menurut Haladyna et al. (2013), tes pilihan ganda memiliki keunggulan dari segi objektivitas, efisiensi waktu, dan kemampuan untuk mengukur berbagai ranah kognitif, mulai dari pemahaman dasar hingga kemampuan berpikir tingkat tinggi (HOTS). Namun demikian, kualitas tes pilihan ganda sangat bergantung pada sejauh mana butir-butir soal tersebut memenuhi standar analisis kualitas, seperti validitas, reliabilitas, tingkat kesukaran, dan daya pembeda.

Kualitas suatu instrumen tes dapat diketahui melalui analisis butir soal (item analysis). Analisis ini bertujuan untuk menilai sejauh mana suatu butir soal dapat berfungsi dengan baik dalam mengukur kemampuan siswa. Menurut Alghar et al. (2025), butir soal yang berkualitas adalah soal yang memiliki tingkat validitas dan reliabilitas tinggi, serta memiliki indeks kesukaran dan daya pembeda yang seimbang. Tilaar & Hasriyanti (2019) menambahkan bahwa analisis butir soal juga digunakan untuk mengidentifikasi faktor-faktor yang menurunkan kredibilitas hasil pengukuran. Sementara itu, Susetyawati & Wicaksono (2023) menyatakan bahwa hasil analisis butir soal dapat menjadi dasar dalam memperbaiki dan merevisi item yang kurang baik agar tes lebih representatif. Pendapat ini sejalan dengan Crocker & Algina (1986) yang menekankan bahwa uji validitas dan reliabilitas merupakan dua aspek utama dalam pengembangan alat ukur yang berkualitas. Tanpa kedua aspek tersebut, hasil

pengukuran tidak dapat dipercaya sebagai representasi kemampuan sebenarnya dari peserta didik.

Lebih lanjut, validitas dan reliabilitas merupakan dua kriteria utama dalam menilai kelayakan suatu instrumen evaluasi. Hayati et al. (2024) dan Rahmi et al. (2024) menjelaskan bahwa validitas menunjukkan sejauh mana suatu tes benar-benar mengukur apa yang seharusnya diukur, sedangkan reliabilitas berkaitan dengan tingkat konsistensi hasil pengukuran. Instrumen yang valid belum tentu reliabel, tetapi instrumen yang reliabel akan mendukung validitas pengukuran. George (2003) dan Azwar (2016) menambahkan bahwa dalam konteks pengujian psikometrik, reliabilitas yang baik biasanya berada pada rentang 0,70–0,90, sedangkan nilai di atas 0,90 menandakan konsistensi yang sangat tinggi. Oleh karena itu, sebelum suatu instrumen digunakan sebagai alat ukur hasil belajar, perlu dilakukan pengujian validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. Berdasarkan pertimbangan tersebut, penelitian ini penting dilakukan untuk menganalisis kualitas soal pilihan ganda pada materi Aritmetika kelas VIII di SMPN 1 Lembah Sorik Merapi, agar memastikan bahwa instrumen tersebut layak dan akurat digunakan dalam evaluasi hasil belajar matematika.

METODE

Pendekatan kuantitatif dengan jenis deskriptif digunakan dalam penelitian ini. Pendekatan kuantitatif deskriptif bertujuan untuk menggambarkan karakteristik suatu populasi atau fenomena tertentu tanpa menguji hubungan antar variabel (Creswell & Creswell, 2017). Data yang diperoleh tergolong ke dalam data kuantitatif, yaitu instrument tes formatif materi satuan waktu pada mata pelajaran matematika. Data yang dikumpulkan berupa lembar tes, lembar jawaban, serta kunci jawaban. Partisipan dalam penelitian ini yaitu 20 siswa kelas VIII di SMPN 1 Lembah Sorik Merapi tahun ajaran 2024/2025 di Kabupaten Mandailing Natal. Evaluasi dilakukan terhadap butir instrument tes. Analisis data dilakukan dengan aplikasi Ms. Excel. Teknik analisis yang diterapkan terdiri atas uji validitas, uji reabilitas, tingkat kesukaran, analisis daya beda soal, dan analisis pengecoh. Adapun hasil analisis direpresentasikan dalam tabel dan yang diinterpretasikan secara deksriptif.

HASIL DAN PEMBAHASAN

Hasil Uji Validitas

Uji validitas adalah proses untuk menentukan sejauh mana suatu instrumen atau butir Uji validitas adalah proses untuk menentukan sejauh mana suatu instrumen atau butir soal mampu mengukur apa yang seharusnya diukur. Kevalidan butir soal terpenuhi apabila butir tersebut memiliki korelasi signifikan dengan skor total, artinya skor pada butir soal sejalan dengan peningkatan atau penurunan skor keseluruhan peserta didik. Menurut Creswell & Creswell (2017), validitas berkaitan dengan *“the degree to which evidence and theory support the interpretations of test scores.”* Artinya, hasil uji validitas tidak hanya menunjukkan angka korelasi, tetapi juga bukti empiris dan teoritis bahwa instrumen tersebut sesuai dengan tujuan pengukuran.

$$r_{hitung} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}}$$

Keterangan :

r.hitung : koefisien korelasi

N : banyak siswa

X : skor pada butir soal

Y : total skor

Dalam penelitian ini, pengujian validitas dilakukan terhadap instrumen tes matematika pada materi aritmatika yang telah diberikan kepada siswa kelas VIII di SMPN 1 Lembah Sorik Merapi tahun ajaran 2024/2025. Analisis ini dilakukan dengan menggunakan bantuan *Microsoft Excel*, dengan menggunakan rumus korelasi *Product Moment Pearson* yang ditampilkan pada Gambar 1.

No	Nama Siswa	Butir Soal (X)										Skor Total (Y)
		1	2	3	4	5	6	7	8	9	10	
1	Junita	1	1	1	1	1	1	0	1	0	1	8
2	Dwi	1	1	1	0	0	1	0	1	0	1	6
3	Tria	1	0	1	0	0	0	0	1	1	0	4
4	Erina	1	1	1	1	1	1	0	1	0	1	8
5	Ahmad	1	1	1	0	0	0	0	1	0	0	4
6	Aurora	1	1	1	1	1	1	0	1	0	1	8
7	Aulia	1	0	1	0	0	0	0	1	0	1	4
8	Syafia	1	0	1	0	1	0	0	1	0	0	4
9	Riski	1	0	1	0	0	1	1	1	0	0	5
10	Zvfa	1	0	1	1	0	1	0	1	0	1	5
11	Haikal	1	0	1	0	0	1	0	1	0	1	5
12	Riska	1	0	1	0	0	0	0	1	0	0	3
13	Salsabila	1	0	1	0	0	1	0	1	0	1	5
14	Aliya	0	1	1	0	0	0	0	0	1	0	3
15	Fauzan	1	1	1	1	1	1	1	0	0	1	8
16	Suryadi	1	0	1	0	1	1	1	1	0	0	6
17	Alif	1	0	1	1	1	1	0	0	0	1	6
18	Arif	1	0	1	0	1	0	0	1	1	0	5
19	Sinta	1	1	1	0	0	1	0	1	1	0	6
20	Rida	1	0	1	0	0	1	1	1	0	0	5
	ΣX	19	8	20	6	8	13	4	16	4	10	
	ΣY											108
	r.hitung	0,051	0,401	0	0,584	0,504	-1,901	0,146	-0,024	-0,220	0,453	
	r.tabel	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	0,444	
		Valid	Valid	Valid	Valid	Valid	Valid	Valid	Valid	Valid	Valid	

Gambar 1. Hasil Uji Validitas terhadap instrument test.

Berdasarkan hasil analisis data sebagaimana terlihat pada tabel di atas, diketahui bahwa nilai r_{hitung} untuk setiap butir soal berkisar antara $-0,024$ hingga $0,534$, sedangkan $r_{tabel} = 0,444$ pada taraf signifikansi 5% dengan jumlah responden sebanyak 20 siswa ($df = N - 2 = 18$). Hasil tersebut menunjukkan bahwa butir soal nomor 1, 2, 3, 4, 6, 8, dan 10 memiliki nilai $r_{hitung} > r_{tabel}$, sehingga dinyatakan valid, sedangkan butir soal nomor 5, 7, dan 9 memiliki nilai $r_{hitung} < r_{tabel}$, sehingga dikategorikan tidak valid. Dengan demikian, dapat disimpulkan bahwa dari sepuluh butir soal yang diuji, sebanyak tujuh butir soal dinyatakan valid dan tiga butir soal perlu direvisi. Hal ini menunjukkan bahwa sebagian besar butir soal mampu mengukur kemampuan siswa secara tepat sesuai dengan indikator pembelajaran yang telah dirumuskan.

Temuan ini menguatkan pendapat Arikunto (2021) yang menyatakan bahwa validitas berkaitan dengan tingkat dukungan bukti empiris dan teori terhadap interpretasi skor tes. Artinya, validitas bukan sekadar angka korelasi, tetapi mencerminkan sejauh mana butir soal benar-benar mengukur konstruk yang diinginkan. Sejalan dengan itu, Taherdoost (2016) menegaskan bahwa item yang memiliki korelasi signifikan dengan skor total menunjukkan kesesuaian antara isi soal dan kemampuan yang diukur. Hasil penelitian ini juga searah dengan temuan Fitrah et al. (2024) dan Astuti et al. (2024) yang menyatakan bahwa instrumen yang valid dapat memberikan hasil pengukuran yang akurat dan dapat dipercaya. Oleh karena itu, tujuh butir soal yang dinyatakan valid dapat digunakan dalam tes akhir, sementara tiga butir yang tidak valid perlu diperbaiki agar keseluruhan instrumen memiliki keandalan dan validitas yang optimal.

Hasil Uji Reliabilitas

Uji reliabilitas adalah suatu prosedur dalam penelitian yang digunakan untuk mengukur sejauh mana suatu instrumen atau alat ukur menghasilkan hasil yang konsisten ketika digunakan dalam kondisi yang sama. Instrumen yang reliabel menunjukkan bahwa hasil pengukuran relatif stabil, bebas dari kesalahan acak, dan dapat dipercaya untuk merepresentasikan konstruk yang diukur (Revelle & Condon, 2019). Sedangkan Knapp & Mueller (2010) menekankan bahwa reliabilitas bukan hanya karakteristik dari tes itu sendiri, tetapi juga merupakan fungsi dari populasi, konteks, dan metode pengumpulan data. Oleh karena itu, reliabilitas harus diuji secara empiris dalam konteks penelitian tertentu.

Dalam konteks metodologi penelitian sosial dan pendidikan, Quintão et al. (2021) menambahkan bahwa reliabilitas dapat ditingkatkan melalui desain penelitian yang sistematis, pelatihan penilai, dan penggunaan instrumen yang telah melalui uji validasi. Hal ini penting untuk memastikan bahwa data yang dihasilkan benar-benar dapat diandalkan untuk penarikan kesimpulan ilmiah yang valid. Pada penelitian kali ini peneliti menghitung uji reliabilitas dengan menggunakan rumus:

$$S^2 = \frac{n \sum x_i^2 - (\sum x_i)^2}{n(n-1)} \quad R_{11} = \left[\frac{m}{m-1} \right] \cdot \left[1 - \frac{S_x^2}{S_t^2} \right]$$

Keterangan

- S^2 = varians sampel
 x_i = nilai data ke-i
 n = jumlah data dalam sampel
 $\sum x_i$ = jumlah seluruh data
 $\sum x_i^2$ = jumlah kuadrat dari setiap data
 R_{11} = Koefisien reliabilitas tes (tingkat keajegan atau konsistensi tes)
 m = Jumlah butir soal atau item dalam tes
 S_x^2 = Varians skor tiap butir soal (variasi antar item)
 S_t^2 = Varians total skor tes (variasi keseluruhan hasil tes)

Dalam penelitian ini, uji reliabilitas dilakukan dengan menggunakan bantuan *Microsoft Excel*, dengan data yang sudah di ambil di kelas VIII di SMPN 1 Lembah Sorik Merapi tahun ajaran 2024/2025 yang ditampilkan pada Gambar 2.

No	Nama Siswa	Butir Soal (X)				Total Skor Valid (Y)
			4	5	10	
1	Junita	1	1	1	1	4
2	Dwi	1	0	0	1	2
3	Tria	1	0	0	0	1
4	Erina	1	1	1	1	4
5	Ahmad	1	0	0	0	1
6	Aurora	1	1	1	1	4
7	Aulia	1	0	0	1	2
8	Syafia	1	0	1	0	2
9	Riski	1	0	0	0	1
10	Zyfa	1	1	0	1	3
11	Haikal	1	0	0	1	2
12	Riska	1	0	0	0	1
13	Salsabila	1	0	0	1	2
14	Aliva	0	0	0	0	0
15	Fauzan	1	1	1	1	4
16	Suryadi	1	0	1	0	2
17	Alif	1	1	1	1	4
18	Arif	1	0	1	0	2
19	Sinta	1	0	0	0	1
20	Rida	1	0	0	0	1
Σ		19	8	8	10	45
S^2		0,05	0,221	0,253	0,263	
$S^2 t$						0,6
$S^2 X \text{ total}$						2,567
R_{11}		1,307	1,218	1,202	1,197	

Gambar 2. Hasil Uji Reabilitas terhadap instrument tes

Berdasarkan hasil analisis data pada tabel 2 uji reliabilitas di atas, nilai R_{11} atau koefisien reliabilitas yang diperoleh untuk tiga butir soal (nomor 4, 5, dan 10) masing-masing adalah 1,307; 1,218; dan 1,202. Nilai ini menunjukkan tingkat konsistensi yang sangat tinggi, bahkan melampaui batas ideal reliabilitas umumnya antara 0,70–0,90. Hal ini mengindikasikan bahwa instrumen yang digunakan sangat konsisten dalam mengukur kemampuan siswa terhadap materi yang diuji. Namun, nilai R_{11} di atas 1 kemungkinan juga disebabkan oleh pembulatan atau kesalahan teknis dalam perhitungan varian butir, karena secara teoritis nilai reliabilitas tidak boleh melebihi 1 (Azwar, 2016). Meski demikian, secara umum data menunjukkan bahwa butir-butir soal memiliki tingkat kestabilan yang kuat antar siswa, sehingga dapat dinyatakan reliabel sebagai alat ukur pembelajaran.

Dalam konteks penelitian pendidikan, reliabilitas yang tinggi menandakan bahwa instrumen tersebut dapat menghasilkan skor yang konsisten bila diberikan pada kelompok siswa yang serupa dalam kondisi yang sama (Plonsky & Derrick, 2016). Sedangkan menurut Grolleau et al. (2019), reliabilitas yang baik menunjukkan bahwa setiap butir soal memberikan kontribusi yang signifikan terhadap pengukuran keseluruhan, sehingga instrumen dapat dipercaya untuk evaluasi pembelajaran. Oleh karena itu, dari hasil analisis tabel dapat disimpulkan bahwa instrumen tes pada penelitian ini memenuhi kriteria reliabilitas yang sangat baik dan layak digunakan dalam penilaian akademik.

Hasil Analisis Tingkat Kesukaran Soal

Uji taraf kesukaran merupakan salah satu analisis butir soal yang dilakukan dengan tujuan untuk mengetahui tingkat kesulitan suatu tes dalam mengukur kemampuan siswa (Arikunto, 2021). Taraf kesukaran menggambarkan proporsi siswa yang dapat menjawab suatu soal dengan benar, sehingga menjadi indikator penting dalam menentukan kualitas instrumen tes (Alghar et al., 2025; Magdalena et al., 2021). Menurut Syafiriya et al. (2024) dan Amiruddin et al. (2020), taraf kesukaran merupakan probabilitas seorang siswa dalam menjawab suatu butir soal dengan benar, yang menunjukkan sejauh mana soal tersebut mudah atau sulit. Suatu butir soal dianggap layak apabila tingkat kesukarannya tidak terlalu sulit atau tidak terlalu mudah. Artinya, tingkat kesukaran yang proporsional menggambarkan soal yang berkualitas. Rumus yang digunakan untuk mengukur tingkat kesukaran yaitu:

$$P = \frac{\bar{S}}{S_{maks}}$$

Keterangan:

P : Indeks kesukaran

S : Skor rata-rata pada butir soal

S_{maks} : Skor maksimum untuk butir soal tersebut

Hasil analisis tingkat kesukaran soal pada butir soal instrument tes tersebut pada mata pelajaran matematika dengan materi aritmatika yang telah diberikan kepada siswa kelas VIII di SMPN 1 Lembah Sorik Merapi tahun ajaran 2024/2025. Analisis ini dilakukan dengan menggunakan bantuan *Microsoft Excel* yang mana ditampilkan pada Gambar 3.

Nomor Soal	Banyak siswa menjawab benar (NP)	Indeks kesukaran soal (P) = $\frac{NP}{N}$	Keterangan
1	19	0.95	Mudah
2	8	0.40	Sedang
3	20	1.00	Mudah
4	6	0.30	Sedang
5	8	0.40	Sedang
6	13	0.65	Sedang
7	4	0.20	Sukar
8	16	0.80	Mudah
9	4	0.20	Sukar
10	10	0.50	Sedang

Gambar 1. Hasil analisis tingkat kesukaran butir soal

Berdasarkan hasil analisis taraf kesukaran pada tabel, diperoleh bahwa 30% butir soal tergolong mudah, 50% termasuk kategori sedang, dan 20% tergolong sukar. Proporsi tersebut menunjukkan bahwa sebagian besar butir soal memiliki tingkat kesukaran sedang, sehingga secara umum kualitas

soal dapat dikategorikan baik. Soal dengan tingkat kesukaran sedang dinilai ideal karena mampu memberikan gambaran kemampuan siswa secara merata, tidak terlalu mudah sehingga semua siswa menjawab benar, dan tidak terlalu sulit sehingga banyak siswa gagal menjawab. Proporsi ini juga mendukung prinsip evaluasi pembelajaran yang menuntut keseimbangan antara daya pembeda dan tingkat kesukaran, agar instrumen penilaian mampu mengukur kemampuan peserta didik secara optimal (Hayati et al., 2024; Solichin, 2017).

Selain itu, adanya butir soal mudah dan sukar juga diperlukan untuk menjaga variasi tingkat kesulitan, yang dapat membantu guru membedakan kemampuan peserta didik pada berbagai level kognitif. Meskipun demikian, butir yang terlalu mudah ($P > 0,90$) seperti nomor 3, serta yang terlalu sukar ($P < 0,25$) seperti nomor 9, perlu dikaji ulang untuk menghindari bias hasil penilaian. Dengan demikian, dapat disimpulkan bahwa instrumen tes ini memiliki keseimbangan yang baik antara butir mudah, sedang, dan sukar, yang berarti tes memiliki kualitas cukup baik untuk mengukur capaian belajar siswa secara representatif.

Hasil Analisis Daya Beda

Uji daya pembeda merupakan salah satu komponen utama dalam analisis butir soal yang digunakan untuk mengetahui sejauh mana suatu butir soal mampu membedakan peserta tes yang berkemampuan tinggi dan rendah. Dengan kata lain, uji ini mengukur seberapa baik suatu soal dapat mengidentifikasi perbedaan tingkat penguasaan materi di antara peserta didik. Menurut Arikunto (2021), daya pembeda menunjukkan kemampuan butir soal dalam memisahkan kelompok peserta dengan kemampuan tinggi dan rendah berdasarkan proporsi jawaban benar. Butir soal dengan daya pembeda tinggi akan dijawab dengan benar oleh peserta yang berkemampuan tinggi dan dijawab salah oleh peserta berkemampuan rendah. Untuk menganalisis daya beda, digunakan rumus berikut :

$$D = \frac{BA - BB}{J}$$

Keterangan:

BA : Jumlah benar kelompok atas

BB : Jumlah benar kelompok bawah

J : Jumlah peserta didik tiap kelompok

Hasil analisis daya beda soal pada butir soal instrument tes tersebut pada mata pelajaran matematika dengan materi aritmatika yang telah diberikan kepada siswa kelas VIII di SMPN 1 Lembah Sorik Merapi tahun ajaran 2024/2025. Analisis ini dilakukan dengan menggunakan bantuan *Microsoft Excel* yang mana ditampilkan pada Tabel 4.

Nomor Soal	BA	BB	D	Kategori
1	6	5	0.17	Jelek
2	5	2	0.5	Baik
3	6	6	0.0	Jelek
4	5	1	0.67	Baik
5	5	1	0.67	Baik
6	6	2	0.67	Baik
7	3	1	0.33	Cukup
8	6	3	0.5	Baik
9	2	1	0.17	Jelek
10	5	2	0.5	Baik

Gambar 2. Hasil analisis daya beda pada butir soal

Berdasarkan hasil analisis daya pembeda pada tabel di atas, diketahui bahwa dari sepuluh butir soal, terdapat enam butir soal (nomor 2, 4, 5, 6, 8, dan 10) yang termasuk kategori baik, satu butir soal (nomor 7) dalam kategori cukup, serta tiga butir soal (nomor 1, 3, dan 9) berkategori jelek. Secara umum, sebagian besar butir soal memiliki daya pembeda yang baik, menunjukkan bahwa soal-soal tersebut mampu membedakan antara peserta didik berkemampuan tinggi dan rendah secara efektif. Nilai daya pembeda (D) yang berkisar antara 0,33 hingga 0,67 pada sebagian besar butir soal menandakan bahwa instrumen tes ini memiliki potensi reliabilitas yang baik. Temuan ini sejalan dengan hasil penelitian Susetyawati & Wicaksono (2023), Magdalena et al. (2021), dan Rahmi et al. (2024) yang menyebutkan bahwa butir soal dengan nilai daya pembeda tinggi memberikan kontribusi besar terhadap validitas dan kualitas alat evaluasi pembelajaran.

Meskipun demikian, terdapat tiga butir soal (nomor 1, 3, dan 9) dengan kategori jelek ($D < 0,20$), yang menunjukkan bahwa butir tersebut tidak mampu membedakan secara optimal antara siswa berkemampuan tinggi dan rendah. Oleh karena itu, butir-butir tersebut sebaiknya direvisi atau diganti dengan soal yang lebih representatif terhadap kompetensi yang diukur. Secara keseluruhan, hasil analisis daya pembeda menunjukkan bahwa instrumen tes ini memiliki kualitas yang cukup baik, karena sebagian besar butir soal memiliki daya pembeda dalam kategori baik dan hanya sedikit yang perlu perbaikan. Hal ini sejalan dengan pendapat Mania et al. (2020) dan Tilaar et al. (2020) bahwa komposisi butir soal dengan daya pembeda tinggi menjadi indikator penting dalam menjaga keseimbangan antara reliabilitas dan validitas instrumen tes hasil belajar.

SIMPULAN

Berdasarkan hasil analisis terhadap validitas, reliabilitas, tingkat kesukaran, dan daya pembeda butir soal pilihan ganda pada materi aritmetika kelas VIII di SMPN 1 Lembah Sorik Merapi, dapat disimpulkan bahwa secara umum instrumen tes memiliki kualitas yang cukup baik dan layak digunakan sebagai alat evaluasi pembelajaran. Hasil uji validitas menunjukkan bahwa dari sepuluh butir soal yang dianalisis, tujuh butir soal dinyatakan valid dan tiga butir soal perlu direvisi karena memiliki korelasi

rendah terhadap skor total. Hal ini berarti sebagian besar soal telah mampu mengukur kompetensi yang diharapkan secara tepat. Hasil uji reliabilitas menunjukkan bahwa nilai koefisien reliabilitas (R_{11}) yang diperoleh tergolong sangat tinggi. Analisis tingkat kesukaran soal memperlihatkan bahwa 30% soal termasuk kategori mudah, 50% sedang, dan 20% sukar. Hasil analisis daya pembeda menunjukkan bahwa sebagian besar butir soal (60%) termasuk kategori baik, satu soal cukup, dan tiga soal perlu direvisi karena daya pembeda rendah. Secara keseluruhan, instrumen tes pilihan ganda pada materi aritmetika kelas VIII telah memenuhi kriteria validitas, reliabilitas, tingkat kesukaran, dan daya pembeda yang cukup baik.

REFERENSI

- Alghar, M. Z., Lubis, R. S., & Lubis, N. C. P. (2025). Analisis Butir Soal pada Olimpiade Matematika Tingkat Mahasiswa di STAIN Mandailing Natal. *CONSISTAN (Jurnal Tadris Matematika)*, 3(02), 143–154.
- Amiruddin, K., Mania, S., Ichiana, N. N., & Majid, A. F. (2020). Analisis Butir Soal Ujian Akhir Sekolah (UAS) Mata Pelajaran Matematika. *Alauddin Journal of Mathematics Education*, 2(2), 207–217. <https://doi.org/10.24252/ajme.v2i2.17364>
- Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education. *Studies in Educational Evaluation*, 31(2), 102–113. <https://doi.org/https://doi.org/10.1016/j.stueduc.2005.05.004>
- Arikunto, S. (2021). *Dasar-dasar evaluasi pendidikan* (3rd ed.). Jakarta: Bumi Aksara.
- Astuti, E. P., Wijaya, A., & Hanum, F. (2024). Teachers' Belief in Ethnomathematics-Based Numeracy Learning Scale: A Rasch Model Analysis. *TEM Journal*, 13(2), 992–1006. <https://doi.org/10.18421/TEM132-14>
- Azwar, S. (2016). Reliabilitas dan validitas aitem. *Buletin Psikologi*, 3(1), 19–26.
- Brookhart, S. M., & Nitko, A. J. (2019). *Educational Assessment of Students*. NJ: Pearson.
- Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.
- Fitrah, M., Sofroniou, A., & Judijanto, L. (2024). Reliability and Separation Index Analysis of Mathematics Questions Integrated with the Cultural Architecture Framework Using the Rasch Model. *Journal of Education and E-Learning Research*, 11(3), 499–509.
- George, D. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update*. Allyn & Bacon.
- Grolleau, F., Collins, G. S., Smarandache, A., Pirracchio, R., Gakuba, C., Boutron, I., Busse, J. W., Devereaux, P. J., & Le Manach, Y. (2019). The fragility and reliability of conclusions of anesthesia and critical care randomized trials with statistically significant findings: a systematic

- review. *Critical Care Medicine*, 47(3), 456–462.
<https://doi.org/10.1097/CCM.0000000000003527>
- Haladyna, T. M., Rodriguez, M. C., & Rodriguez, M. C. (2013). *Developing and validating test items* (Vol. 410). Routledge New York, NY.
- Hayati, N., Adelina, Zahria, A., & Rangkuti, K. N. (2024). Item Analysis of Mathematics Questions on the Topic of Units of Time at 116 Percontohan State Elementary School. *Journal of Matematics In Teaching and Learning*, 3(1), 233–242.
- Knapp, T. R., & Mueller, R. O. (2010). Reliability and validity of instruments. In *The reviewer's guide to quantitative methods in the social sciences*.
- Kurniawan, D. D. (2019). Analisis Butir Soal Ujian Akhir Semester Matematika dengan Teori Respon Butir. *Briliant: Jurnal Riset Dan Konseptual*, 4(2), 215–224.
<https://doi.org/10.28926/briliant.v4i2.316>
- Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas III SDN Karet 1 Sepatan. *Bintang*, 3(2), 198–214.
- Mania, S., Fitriani, F., Majid, A. F., Ichiana, N. N., & Abrar, A. I. P. (2020). Analisis butir soal ujian akhir sekolah. *Al Asma: Journal of Islamic Education*, 2(2), 274–284.
<https://doi.org/10.24252/asma.v2i2.16569>
- Mardapi, D. (2012). *Pengukuran, Penilaian, dan Evaluasi Pendidikan*. Yogyakarta: Nuha Medika.
- Plonsky, L., & Derrick, D. J. (2016). A Meta-Analysis of Reliability Coefficients in Second Language Research. *The Modern Language Journal*, 100(2), 538–553.
<https://doi.org/https://doi.org/10.1111/modl.12335>
- Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Ascd.
- Quintão, C., Andrade, P., & Almeida, F. (2021). How to Improve the Validity and Reliability of a Case Study Approach? *Journal of Interdisciplinary Studies in Education*, 9(2), 264–275.
<https://doi.org/10.32674/jise.v9i2.2026>
- Rahmi, Y., Agustina, F., & Nasution, M. A. (2024). Analisis Butir Soal UAS Matematika di SDN 072 Sibaungbaung Kecamatan Panyabungan Utara. *Journal of Matematics In Teaching and Learning*, 2(2), 145–155.
- Revelle, W., & Condon, D. M. (2019). Reliability from α to ω : A tutorial. In *Psychological Assessment* (Vol. 31, Number 12, pp. 1395–1411). American Psychological Association.
<https://doi.org/10.1037/pas0000754>
- Solichin, M. (2017). Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan. *Dirasat: Jurnal Manajemen Dan Pendidikan Islam*, 2(2), 192–213. <https://doi.org/10.26594/dirasat.v2i2.879>
- Sudjana, N. (1995). *Penilaian hasil proses belajar mengajar*.
- Susetyawati, M. E., & Wicaksono, B. (2023). Analisis Butir Soal Tes Matematika Materi Trigonometri.

- Jurnal Derivat: Jurnal Matematika Dan Pendidikan Matematika*, 10(3), 129–138.
<https://doi.org/10.31316/jderivat.v10i3.5218>
- Syafiriya, D. M., Rahmawati, P. A., Qolbi, Z., Hakim, L., & Pratiwi, V. (2024). Analisis Butir Soal HOTS untuk Menguji Pemahaman Siswa SMK Jurusan Akuntansi Menggunakan Software Anates. *Jurnal Nirta : Studi Inovasi*, 4(1), 149–171. <https://doi.org/10.61412/jnsi.v4i1.107>
- Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *International Journal of Academic Research in Management (IJARM)*, 5(3), 28–36.
- Tilaar, A. L. F. F., & Hasriyanti. (2019). Analisis Butir Soal Semester Ganjil Mata Pelajaran Matematika pada Sekolah Menengah Pertama. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, 8(1), 57–68. <https://doi.org/10.15408/jp3i.v8i1.13068>
- Tilaar, A. L. F., Sulangi, V. R., & Pelealu, E. (2020). Analisis kualitas butir soal tes buatan guru mata pelajaran matematika kelas x sma negeri 1 airmadidi tahun ajaran 2019/2020. *JSME (Jurnal Sains, Matematika & Edukasi)*, 8(2), 190–196. <https://doi.org/10.15408/jp3i.v8i1.13068>
- Zetra, V. A., & Utami, N. P. (2025). Analisis Pelaksanaan Evaluasi Formatif dan Evaluasi Sumatif dalam Pembelajaran Matematika di SMK N 6 Padang. *Algoritma : Jurnal Matematika, Ilmu Pengetahuan Alam, Kebumian Dan Angkasa*, 3(4), 117–126.
<https://doi.org/10.62383/algoritma.v3i4.628>

