https://journal.nabest.id/index.php/annajah
e-ISSN: 2964-965X (Media Online) | p-ISSN: 2964-9633 (Media Cetak)
Vol. 5 No.4 (Juli 2026)

Submitted: 25-05-2026 | Accepted: 04-06-2026 | Published: 01-07-2026

ANALISIS KUALITAS BUTIR SOAL ASESMEN SUMATIF AKHIR TAHUN (ASAT) MATA PELAJARAN AL-QUR'AN HADITS DALAM KURIKULUM MERDEKA DI MTS AL-MASRURIYAH BATURRADEN

Nasrul Taufikurrohman[1], Donny Khoirul Aziz[2]
[1,2] UIN Prof. K.H. Saifuddin Zuhri Purwokerto, Indonesia
Email: [1]nasrulrroman@gmail.com [2]dony@uinsaizu.a.id

Abstract

The End-of-Year Summative Assessment (ASAT) under the Merdeka Curriculum is designed to formally evaluate student learning progression. This descriptive quantitative study assesses the psychometric quality of Grade IX ASAT multiple-choice items for the Al-Qur'an Hadith subject at MTs Al-Masruriyah Baturraden (academic year 2025/2026), compiled by KKM Pelangi. Grounded in Classical Test Theory (CTT) and analyzed via ANATES V4, the study evaluated the entire student population. The findings revealed a discrepancy between theoretical design and actual field performance. While the assessment tool demonstrated perfect alignment in content validity with a proportional difficulty distribution dominated by moderate items, empirical analysis showed that nearly half of the total items were statistically invalid. This empirical limitation corresponds with the instrument's moderate internal consistency (reliability) in the "sufficient" category, driven by nonfunctioning distractors in the majority of items and critical discrimination index anomalies, such as negative discrimination on specific items. These technical limitations of the distractors systematically increased the guessing factor. Consequently, this study recommends reconstructing the incorrect options based on student misconceptions, adjusting the answer keys, and introducing additional homogeneous items to enhance the overall precision and reliability of the evaluation instrument.

Keywords: *Educational Evaluation, Summative Assessment, Classical Test Theory, Qur'an Hadith, ANATES V4*

Abstrak

Asesmen Sumatif Akhir Tahun (ASAT) dalam Kurikulum Merdeka bertujuan mengukur ketercapaian Capaian Pembelajaran secara formal. Penelitian deskriptif kuantitatif ini mengevaluasi kualitas psikometrik butir soal pilihan ganda ASAT mata pelajaran Al-Qur'an Hadits Kelas IX di MTs Al-Masruriyah Baturraden tahun ajaran 2025/2026 yang disusun kolaboratif bersama KKM Pelangi. Berdasarkan Teori Tes Klasik (CTT) dengan bantuan program ANATES V4, analisis dilakukan terhadap respons seluruh populasi siswa. Hasil penelitian menunjukkan adanya kesenjangan antara aspek teoretis dan empiris. Validitas isi naskah soal mencapai keselarasan mutlak dengan sebaran tingkat kesukaran yang proporsional didominasi kategori sedang. Namun secara empiris, hampir setengah dari total butir soal terbukti tidak valid secara statistik. Kelemahan empiris ini sejalan dengan derajat konsistensi internal (reliabilitas) instrumen yang berada pada kategori cukup, lemahnya fungsi opsi pengecoh yang mayoritas tidak bekerja secara efektif, serta ditemukannya anomali berupa daya pembeda negatif pada butir soal tertentu. Lemahnya distraktor secara sistematis meningkatkan faktor tebakan acak (guessing factor). Berdasarkan temuan ini, direkomendasikan melakukan rekonstruksi opsi pengecoh berbasis pola miskonsepsi siswa, penyesuaian kunci jawaban, serta penambahan butir soal homogen guna meningkatkan reliabilitas instrumen evaluasi hasil belajar madrasah.

Kata Kunci: Evaluasi Pendidikan, Asesmen Sumatif Akhir Tahun, Teori Tes Klasik, Al-Qur'an Hadits, ANATES V4

A. PENDAHULUAN

Pendidikan pada dasarnya merupakan upaya nyata dan terencana untuk membantu setiap peserta didik dalam menemukan serta mengasah potensi terbaik di dalam dirinya, guna membangun spiritualitas keagamaan, akhlak mulia, dan kecerdasan yang berintegritas (Pristiwanti et al., 2022). Dalam proses instruksional, efektivitas pencapaian hasil belajar sangat bergantung pada sinergi berbagai komponen kunci, mulai dari perencanaan, pelaksanaan, hingga evaluasi akhir (Hamzah, 2022). Perubahan paradigma pendidikan melalui Kurikulum Merdeka menuntut pergeseran mendasar dalam cara pendidik memaknai proses penilaian, di mana asesmen diposisikan sebagai satu kesatuan utuh dengan pembelajaran (*assessment as/for learning*) untuk mengevaluasi dan memperbaiki kualitas belajar secara berkelanjutan (Mulyono & Sulistyani, 2022).

Demi menghasilkan data evaluasi yang kredibel, instrumen penilaian wajib memenuhi kriteria kualitas psikometrik yang meliputi dimensi validitas, reliabilitas, objektivitas, dan keadilan (Nitko & Brookhart, 2006). Hal ini sangat esensial pada mata pelajaran Al-Qur'an Hadits, yang karakteristiknya tidak sekadar menguji dimensi kognitif teoretis atau kapasitas ingatan (*Lower Order Thinking Skills*), melainkan diarahkan pada pembentukan karakter spiritual, moral, dan internalisasi nilai keagamaan (Fatmawati, 2025). Oleh karena itu, Asesmen Sumatif Akhir Tahun (ASAT) Al-Qur'an Hadits harus melalui perencanaan psikometrik yang cermat guna memotret pemahaman mendalam dan nalar kritis siswa secara komprehensif (Arifin, 2009).

Namun demikian, fenomena empiris di MTs Al-Masruriyah Baturraden menunjukkan adanya kesenjangan operasional. Meskipun naskah soal ASAT Al-Qur'an Hadits Kelas IX disusun secara kolaboratif melalui Kelompok Kerja Madrasah (KKM) Pelangi guna menjamin standar kualitas kurikulum, instrumen tersebut belum ditindaklanjuti dengan analisis butir soal pasca-ujian secara ilmiah. Ketiadaan quality control psikometrik ini berimplikasi nyata pada pelaksanaan ujian tanggal 26 Februari 2026, di mana mayoritas siswa mengalami disorientasi kognitif akibat redaksi soal yang ambigu, sehingga perolehan nilai rata-rata siswa tidak mampu melampaui Kriteria Ketercapaian Tujuan Pembelajaran (KKTP).

Penggunaan instrumen evaluasi yang tidak teruji kualitasnya secara mendalam berisiko melahirkan data capaian belajar yang bias, menyesatkan, dan menghambat perbaikan pembelajaran (Sa'adah et al., 2025). Berdasarkan urgensi tersebut, penelitian ini bertujuan untuk membedah secara objektif kualitas psikometrik dari 20 butir soal pilihan ganda ASAT Al-Qur'an Hadits Kelas IX di MTs Al-Masruriyah Baturraden. Analisis dilakukan berdasarkan paradigma Teori Tes Klasik (Classical Test Theory atau CTT) berbantuan program ANATES versi 4 guna menguji tingkat validitas, reliabilitas, tingkat kesukaran, daya pembeda, serta efektivitas pengecoh secara presisi, cepat, dan bebas dari risiko kesalahan hitung manual (*human error*) (Wiguna, 2021).

B. METODE

Penelitian ini menerapkan jenis penelitian kuantitatif dengan pendekatan deskriptif kuantitatif, bertujuan memaparkan secara sistematis karakteristik psikometrik dari butir soal ASAT Al-Qur'an Hadits tanpa melakukan pengujian hipotesis (Sugiyono, 2013). Penelitian dilaksanakan di MTs Al-Masruriyah Baturraden, Kabupaten Banyumas, Jawa Tengah, sepanjang bulan Februari hingga April tahun ajaran 2025/2026. Subjek penelitian ini adalah seluruh siswa kelas IX MTs Al-Masruriyah Baturraden yang berjumlah 69 siswa, terdiri dari Kelas IX-A (25 siswa), Kelas IX-B (23 siswa), dan Kelas IX-C (21 siswa). Mengingat seluruh anggota populasi dilibatkan secara utuh.

Instrumen yang dianalisis berupa dokumen asli naskah soal pilihan ganda ASAT Al-Qur'an Hadits semester genap tahun ajaran 2025/2026 sebanyak 20 butir soal, kunci jawaban, dan lembar jawaban dari 68 siswa (Masitoh, 2022). Teknik pengumpulan data dilakukan melalui triangulasi metode, yaitu observasi nonsistematis terhadap pelaksanaan ujian, wawancara tidak terstruktur dengan kepala madrasah dan guru mata pelajaran untuk memperoleh data awal, serta dokumentasi berkas evaluasi (Sugiyono, 2013).

Analisis butir soal dilakukan menggunakan pendekatan Teori Tes Klasik berbantuan perangkat lunak ANATES versi 4.0.9 dengan parameter perhitungan sebagai berikut:

1. Validitas empiris dihitung menggunakan koefisien korelasi Product Moment Pearson antara skor butir (X) dengan skor total (Y) (Arifin, 2009):

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

Keterangan:
r_{xy} : Koefisien validitas
N : Banyaknya subyek
X : Jumlah yang menjawab benar tiap soal
Y : Jumlah total nilai yang menjawab benar tiap soal
Butir soal dinyatakan valid secara empiris apabila nilai $r_{xy} > r_{tabel}$ pada taraf signifikansi α=0,05 untuk uji dua arah (df=66, sehingga r_{tabel}=0,2387).

2. Reliabilitas instrumen diestimasi dengan formula koefisien Cronbach's Alpha (α) guna mengukur konsistensi internal. Tingkat kesukaran soal dihitung dengan rumus (Arifin, 2009):

$$a = \frac{R}{R-1}\left(1 - \frac{\sum a_i^2}{a_x^2}\right)$$

Keterangan:
a : Koefisien reliabilitas
R : Jumlah butir soal
$a\frac{2}{i}$: Varian skor butir soal
$a\frac{2}{x}$: Varian total skor

3. Tingkat kesukaran (P) dihitung untuk mengetahui proporsi siswa yang menjawab benar, dengan klasifikasi: 0,00 - 0,30 (Sukar), 0,31 - 0,70 (Sedang), dan 0,71 - 1,00 (Mudah). Tingkat kesukaran soal dihitung dengan rumus (Sudijono, 2013):

$$P = \frac{B}{JS}$$

Keterangan:
P : Indeks Kesukaran butir
B : Banyak siswa yang menjawab benar
N : Jumlah seluruh peserta tes

4. Indeks daya pembeda (DP) dianalisis untuk memisahkan kelompok siswa berkemampuan tinggi (upper group) dan berkemampuan rendah (lower group). Untuk menghitung daya pembeda setiap butir soal dapat digunakan rumus sebagai berikut (Arifin, 2009):

$$DP = \frac{(WL - WH)}{n}$$

Keterangan :
DP : Daya Pembeda
WL : Jumlah pesetta didik yang gagal dari kelompok bawah.
WH : Jumlah peserta didik yang gagal dari kelompok atas.
n : 27 % x N

5. Efektivitas pengecoh (IP) diuji untuk memastikan pilihan jawaban salah dipilih minimal oleh 5% dari total populasi di luar kunci jawaban. Efektivitas pengecoh dapat diukur menggunakan rumus (Riani et al., 2020):

$$IP = \frac{P}{(N-B)/(n-1)} \times 100\%$$

Keterangan :
IP : Indeks pengecoh
P : Jumlah peserta didik yang memilih pengecoh
N : Jumlah peserta didik yang mengikuti tes
B : Jumlah peserta didik yang menjawab benar pada setiap soal
N : Jumlah bilangan alternatif jawaban
1 : Bilangan tetap

C. HASIL DAN PEMBAHASAN

1. Validitas Isi dan Empiris

Berdasarkan analisis teoretis terhadap kisi-kisi dan silabus, instrumen ASAT Al-Qur'an Hadits Kelas IX menunjukkan tingkat validitas isi yang sempurna sebesar 100%. Seluruh 20 butir soal pilihan ganda dirancang selaras dengan Kompetensi Dasar (KD) semester genap. Pemetaan keselarasan validitas isi tersaji pada Tabel 1:

Tabel 1. Hasil Pemetaan Validitas Isi Instrumen ASAT

Nomor Soal	Kompetensi Dasar / Lingkup Materi Kurikulum	Status Keselarasan	Keterangan
1 - 8	KD 3.4 Ketentuan bacaan gharib (Imalah, Isymam, Tashil, Naql, Mad/Qashr) dalam Al-Qur'an	Sesuai	Valid
9 - 17	KD 3.5 Kandungan Q.S. Abasa (80): 1-10 dan Q.S. Al-Mujadalah (58): 11 tentang menuntut ilmu	Sesuai	Valid
18 - 20	KD 3.6 Kandungan H.R. Muslim dari Abu Hurairah dan H.R. Ibnu Majah dari Safwan bin Assal Al-Muradi	Sesuai	Valid

Namun demikian, hasil pengujian validitas empiris menggunakan program ANATES V4 menunjukkan deviasi yang signifikan (Wiguna, 2021). Melalui nilai batas kritis r_{tabel}=0,2387 untuk N=68, output koefisien korelasi (r_{xy}) masing-masing butir soal dipaparkan pada Tabel 2:

Tabel 2. Hasil Output Validitas Empiris Butir Soal ASAT

No. Soal	Korelasi (r_{xy})	r_{tabel} (N=69)	Status Signifikansi	Klasifikasi Empiris
1	0,149	0.2387	Tidak Signifikan	**Tidak Valid**
2	0,394	0.2387	Signifikan	**Valid**
3	-0,128	0.2387	Tidak Signifikan	**Tidak Valid**
4	0,464	0.2387	Signifikan	**Valid**
5	0,239	0.2387	Signifikan	**Valid**
6	0,573	0.2387	Signifikan	**Valid**
7	0,141	0.2387	Tidak Signifikan	**Tidak Valid**
8	0,687	0.2387	Signifikan	**Valid**
9	0,518	0.2387	Signifikan	**Valid**
10	0,302	0.2387	Signifikan	**Valid**
11	0,417	0.2387	Signifikan	**Valid**
12	0,024	0.2387	Tidak Signifikan	**Tidak Valid**
13	0,183	0.2387	Tidak Signifikan	**Tidak Valid**
14	-0,009	0.2387	Tidak Signifikan	**Tidak Valid**
15	0,048	0.2387	Tidak Signifikan	**Tidak Valid**
16	0,482	0.2387	Signifikan	**Valid**
17	0,250	0.2387	Signifikan	**Valid**
18	0,191	0.2387	Tidak Signifikan	**Tidak Valid**
19	0,579	0.2387	Signifikan	**Valid**
20	0,181	0.2387	Tidak Signifikan	**Tidak Valid**

Nilai korelasi hitung (r_{xy}) tersebut selanjutnya diperbandingkan dengan nilai r_{tabel} pada taraf signifikansi $\alpha = 0{,}05$ untuk uji dua arah. Melalui jumlah subjek N = 68 (df = 68 - 2 = 66), didapatkan nilai r_{tabel} sebesar 0.2387. Suatu butir soal dinyatakan valid secara empiris apabila nilai koefisien korelasi hitung lebih besar daripada r_{tabel} ($r_{xy} > 0.2387$). Sebaliknya, jika $r_{xy} \leq 0.2387$, butir soal tersebut diklasifikasikan sebagai tidak valid (Magdalena et al., 2021).

Berdasarkan data di atas, dari 20 butir soal yang dianalisis, hanya terdapat 11 butir soal (55%) yang dinyatakan valid secara empiris, sedangkan 9 butir soal sisanya (45%) tidak valid. Rata-rata koefisien validitas tercatat pada kategori rendah yaitu r_{xy}=0,37. Deviasi kualitatif ini mengindikasikan adanya inkonsistensi pengukuran yang tajam antara rencana teoretis dengan daya ukur riil instrumen di lapangan (Akhmadi, 2021). Inkonsistensi ini dipicu oleh konstruksi redaksi soal yang membingungkan atau terlalu mengarah pada ingatan hafalan belaka.

2. Reliabilitas Tes

Berdasarkan output pengolahan data menggunakan program ANATES V4, diperoleh koefisien reliabilitas konsistensi internal Cronbach's Alpha sebesar 0,54. Rata-rata perolehan skor siswa tercatat sebesar 9,90 dengan simpang baku sebesar 2,60 dari total skor maksimal

teoretis sebesar 20. Menurut kriteria interpretasi koefisien reliabilitas dari Guilford, nilai koefisien 0,54 berada dalam rentang 0,40 - 0,60, yang berarti memiliki kategori "Cukup" (Arifin, 2009). Derajat keajegan ini membuktikan bahwa instrumen tersebut memiliki keandalan moderat untuk dioperasionalkan di tingkat madrasah, namun masih rentan terdistorsi oleh varians kesalahan pengukuran (*standard error of measurement*). Untuk meningkatkan koefisien reliabilitas ke tingkat yang tinggi (≥ 0,70), disarankan untuk melakukan *item pruning* dengan mengeliminasi soal tidak valid dan merekonstruksi opsi pengecoh.

3. Tingkat Kesukaran

Tingkat kesukaran menunjukkan seberapa mudah atau susah suatu butir soal bagi siswa. Ujian yang ideal harus memiliki komposisi yang proporsional, artinya didominasi oleh soal-soal kategori sedang. Soal yang kelewat mudah tidak akan mampu merangsang daya pikir kritis siswa, sedangkan soal dengan tingkat kesulitan yang melampaui batas kognitif rata-rata siswa bisa memadamkan antusiasme belajar siswa karena dianggap mustahil untuk dikerjakan (Sudijono, 2013). Distribusi tingkat kesukaran dari 20 butir soal pilihan ganda ASAT Al-Qur'an Hadits disajikan pada Tabel 3:

Tabel 3. Distribusi Tingkat Kesukaran Butir Soal ASAT

Klasifikasi Kesukaran	Rentang Indeks (P)	Nomor Butir Soal	Jumlah Butir	Persentase
Sukar	0,00 - 0,30	3, 9, 13, 14, 15	5	25%
Sedang	0,31 - 0,70	1, 2, 5, 6, 7, 8, 12, 16, 18, 19, 20	11	55 %
Mudah	0,71 - 1,00	4, 10, 11, 17	4	20%

Hasil analisis tingkat kesukaran pada 20 butir soal pilihan ganda ASAT Al-Qur'an Hadits menunjukkan sebaran sebagai berikut: sebanyak 5 butir soal (25%) dalam kategori kesulitan tinggi (Sukar), 12 butir soal (55%)pada taraf kesulitan moderat (Sedang), dan 4 butir soal (20%) dalam kriteria tingkat kesulitan rendah (Mudah). Dominasi soal sedang (55%) pada instrumen Al-Qur'an Hadits di MTs Al-Masruriyah Baturraden membuktikan bahwa beban tes ini telah selaras dengan kapasitas kognitif rata-rata peserta didik di madrasah tersebut.

4. Daya Pembeda

Daya pembeda mengevaluasi sensitivitas suatu butir pertanyaan dalam memisahkan kelompok siswa berkemampuan tinggi *(upper group)* dan berkemampuan rendah *(lower group)* (Arikunto, 2021). Sebaran indeks daya pembeda dari hasil analisis disajikan pada Tabel 4:

Tabel 4. Distribusi Daya Pembeda Butir Soal ASAT

Kriteria Daya Pembeda	Rentang Indeks (DP)	Nomor Butir Soal	Jumlah Butir	Persentase
Tidak Baik / Negatif	< 0,00	3	1	5%
Jelek (Poor)	0,00 - 0,20	1, 7, 11, 12, 13, 14, 15, 20	7	35%
Cukup (Satisfactory)	0,21 - 0,40	2, 4, 5, 10, 11, 17, 18	7	35%
Baik (Good)	0,41 - 0,70	9, 16	2	10%
Baik Sekali (Excellent)	0,71 - 1,00	6, 8, 19	3	15%

Berdasarkan Tabel 4, sebanyak 7 butir soal (35%) dikategorikan jelek, 7 butir soal (35%) dikategorikan cukup, 2 butir soal (10%) dikategorikan baik, 3 butir soal (15%) dikategorikan baik sekali, dan 1 butir soal (5%) dikategorikan tidak baik. Merujuk pada sebaran parameter psikometrik tersebut, mayoritas instrumen tes didominasi oleh butir dengan indeks diskriminasi pada level cukup (*satisfactory*) dengan presentase (35%) dan jelek (*poor*) dengan presentase (35%). Namun,

masih ada beberapa soal yang perlu diperbaiki unutk meningkatkan kualitas tes secara keseluruhan. Pendidik disarankan untuk mempertahankan butir-butir soal berkategori "Baik" dan "Baik Sekali" sebagai standar emas dalam bank soal madrasah.

5. Efektivitas Pengecoh

Analisis pengecoh bertujuan untuk mengevaluasi seberapa fungsional opsi jawaban yang tidak benar dalam menjalankan perannya sebagai pengalih perhatian bagi siswa yang belum menguasai kompetensi. Opsi pengecoh (*distraktor*) merupakan alternatif pilihan jawaban di luar kunci jawaban yang dirancang secara sistematis untuk mengalihkan konsentrasi peserta didik yang belum menguasai materi pokok yang sedang diujikan. Konstruksi setiap pilihan jawaban salah ini wajib dirancang sedemikian rupa sehingga mempunyai daya pikat yang kuat bagi peserta didik dalam menentukan keputusan jawaban pada butir soal (Widari, 2021). Distribusi keberfungsian pengecoh disajikan pada Tabel 5:

Tabel 5. Klasifikasi Efektivitas Pengecoh (Distraktor)

Banyak Pengecoh Berfungsi	Kategori Efektivitas	Nomor Butir Soal	Jumlah Butir	Persentase
3	Baik	3	1	5%
2	Cukup	2, 6, 7, 13, 19, 20	6	30%
1	Kurang Baik	1, 4, 5, 8, 9, 10, 11, 14, 16, 18	10	50%
0	Sangat Tidak Baik	12, 15, 17	3	15%

Kualitas pilihan jawaban pengecoh dari 20 butir soal ujian ini bisa dibilang masih kurang maksimal. Tercatat hanya 1 soal (5%) yang berkategori baik dengan tiga pengecoh yang efektif. Selanjutnya, ada 6 soal (30%) berkategori cukup karena memiliki dua fungsi opsi jawaban yang berjalan baik. Sementara itu, separuh dari total soal (50% atau 10 butir) berkategori kurang baik karena hanya satu pengecohnya yang berfungsi, dan 3 soal sisanya (15%) masuk kelompok sangat tidak baik. Opsi distraktor gagal menarik perhatian minimal 5% dari total responden di luar kunci jawaban. Struktur pengecoh yang lemah ini berimplikasi langsung pada tingginya faktor tebakan (*guessing factor*). Jika pilihan jawaban tidak benar dirancang secara tidak homogen, tidak logis, atau terlalu kontras dengan kunci jawaban, siswa berkemampuan rendah dapat mengeliminasi opsi salah tersebut secara instan tanpa melakukan penalaran mendalam. Ini penyebab utama rendahnya indeks validitas empiris dan reliabilitas instrumen di madrasah ini.

D. PENUTUP

Kesimpulan

Berdasarkan pengujian psikometrik Teori Tes Klasik berbantuan software ANATES V4 terhadap instrumen pilihan ganda ASAT Al-Qur'an Hadits Kelas IX di MTs Al-Masruriyah Baturraden, ditarik kesimpulan sebagai berikut:

1. Validitas: Terdapat kesenjangan metodologis yang lebar antara tingkat validitas isi yang sempurna (100%) dengan tingkat validitas empiris yang tergolong rendah dengan korelasi rata-rata r_{xy}=0,39. Dari 20 butir soal, hanya terdapat 11 butir soal (55%) yang valid secara empiris, sedangkan 9 butir soal (45%) dinyatakan tidak valid karena rumusan kalimat stimulus yang ambigu.
2. Reliabilitas: Koefisien reliabilitas konsistensi internal instrumen sebesar 0,54 diklasifikasikan ke dalam kategori "Cukup" (moderat). Nilai ini menunjukkan bahwa keajegan instrumen masih rentan terdistorsi oleh varians kesalahan pengukuran luar, akibat tingginya faktor tebakan.
3. Tingkat Kesukaran: Sebaran indeks kesukaran telah memenuhi prinsip proporsionalitas evaluasi yang ideal, didominasi oleh butir soal berkategori sedang sebesar 55% (11 butir), diikuti soal sukar sebesar 25% (5 butir), dan kategori mudah 20% (4 butir).

4. Daya Pembeda: Sensitivitas daya pembeda instrumen dinilai fungsional dengan dominasi kategori cukup sebesar 35% (7 butir), baik sekali 15% (3 butir), baik 10% (2 butir), jelek 35% (7 butir), dan kategori tidak baik/negatif sebesar 5% (1 butir).
5. Efektivitas Pengecoh: Keberfungsian opsi distraktor teridentifikasi sebagai kelemahan teoretis terbesar instrumen ini, di mana mayoritas butir soal (65%) berada pada kategori kurang baik (50%) atau sangat tidak baik (15%). Lemahnya distraktor memicu tingginya faktor tebakan.

Keterbatasan Penelitian

Penelitian ini membatasi fokus kajian psikometrik hanya pada butir soal objektif berformat pilihan ganda sebanyak 20 butir, sehingga belum memotret kualitas format soal menjodohkan, benar-salah, atau esai yang diujikan dalam draf ASAT secara keseluruhan. Selain itu, karena pengujian didasarkan pada model *Teori Tes Klasik* (CTT), parameter psikometrik yang dihasilkan masih bersifat bergantung pada karakteristik sampel (*sample dependency*), sehingga rentan mengalami fluktuasi jika diujikan pada kelompok siswa yang berbeda.

Saran

Beberapa saran praktis dirumuskan guna meningkatkan kualitas sistem penilaian madrasah:

1. Bagi Pendidik dan Penyusun KKM Pelangi: Sangat disarankan untuk melakukan perbaikan total terhadap 9 butir soal tidak valid dan 6 butir soal yang memiliki daya pembeda jelek/negatif. Konstruksi opsi distraktor harus didasarkan pada pola miskonsepsi siswa yang homogen dan logis. Sebaliknya, butir soal yang terbukti memiliki kualitas psikometrik tinggi layak disimpan secara terstruktur dalam bank soal madrasah.
2. Bagi Manajemen Madrasah: Sekolah perlu menyelenggarakan program peningkatan kompetensi pedagogis guru secara berkala, khususnya pelatihan teknis analisis butir soal berbasis digital dengan bantuan program ANATES V4.
3. Bagi Peneliti Selanjutnya: Kajian evaluasi ini dapat dikembangkan di masa depan dengan memadukan pendekatan Teori Tes Klasik dengan Teori Respon Butir guna meminimalisasi bias ketergantungan sampel.

DAFTAR PUSTAKA

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

Arifin, Z. (2009). *Evaluasi Pembelajaran*. Bandung: Remaja Rosdakarya.

Arikunto, S. (2021). *Dasar-Dasar Evaluasi Pendidikan Edisi 3*. Bumi Aksara.

Fatmawati, E. (2025). Reconceptualizing Assessment In Islamic Education: A Critical Review Of Madrasah Evaluation Practices In The 21st Century. *Journal Of Quality Assurance In Islamic Education (Jqaie)*, *5*(2), 109–119.

Hamzah, H. (2022). *Strategi Pembelajaran Guru Edukatif*. Cv. Azka Pustaka.

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.

Masitoh, M. N. (2022). *Analisis Butir Soal Penilaian Tengah Semester Genap Mata Pelajaran Pendidikan Agama Islam Kelas V Di Sdn 1 Bumiharjo Tahun Ajaran 2021/2022*. Institut Agama Islam Nahdlatul Ulama (IAINU) Kebumen.

Mulyono, R., & Sulistyani, F. (2022). Implementasi Kurikulum Merdeka (IKM) Sebagai Sebuah Pilihan Bagi Satuan Pendidikan: Kajian Pustaka. *Didaktik: Jurnal Ilmiah Pgsd Stkip Subang*, *8*(2), 1999–2019.

Nitko, A. J., & Brookhart, S. M. (2006). *Educational Assessment Of Students*. Prentice-Hall, Inc.

Pristiwanti, D., Badariah, B., Hidayat, S., & Dewi, R. S. (2022). *Pengertian Pendidikan.*

Jurnalpendidikandankonseling (Jpdk), 4 (6), 7911–7915.

Riani, D., Almujab, S., Dina, A., Fitriani, F., & Budiarto, R. (2020). Analisis Butir Soal Dan Kemampuan Siswa Dalam Menjawab Soal Ujian Nasional Pada Mata Pelajaran Ekonomi. *Oikos: Jurnal Kajian Pendidikan Ekonomi Dan Ilmu Ekonomi*, *4*(1), 70–79.

Sa'adah, L., Fakhruddin, A., & Anwar, S. (2025). Assessment For Learning And Value Internalization: Cognitive Assessment In Islamic Religious Education At Indonesian Middle Schools. *Bulletin Of Indonesian Islamic Studies*, *4*(2), 935–960.

Sudijono, A. (2013). *Pengantar Evaluasi Pendidikan*.

Sugiyono. (2013). *Metode Penelitian Kuantitatif, Kualitatif, Dan R & D*. Alfabeta.

Widari, N. A. (2021). Kualitas Butir Soal Pilihan Ganda Ujian Akhir Semester Genap Bahasa Indonesia Kelas X Buatan Mahasiswa Ditinjau Dari Segi Taraf Kesukaran. *Daya Pembeda, Dan Efektivitas Pengecoh*, 1–16.

Wiguna, S. (2021). *Withdrawn: Aplikasi Anates Dalam Evaluasi Pembelajaran*.