AN-NAHDLOH
Journal of Education and Islamic Studies

AN-NAHDLOH: *Journal of Education and Islamic Studies*
Volume 1, Number 4, September 2026; E-ISSN: 3109-8614
Available online at: https://ejournal.magisterpaiunira.com/index.php/annahdloh

Received: May 2026	Accepted: May 2026	Published: September 2026

Evaluasi Kualitas Butir Soal Asesmen Sumatif Pendidikan Agama Islam Kelas VIII di SMP Negeri 2 Kaligondang (Analisis Validitas, Reliabilitas, Tingkat Kesukaran, Daya Beda, dan Efektivitas Pengecoh)

Umul Habibah, Maria Ulpah
Universitas Islam Negeri Profesor KH. Saifuddin Zuhri Purwokerto, Universitas Islam Negeri Profesor KH. Saifuddin Zuhri Purwokerto
Email: umulhabibbah@gmail.com, maria_1511@uinsaizu.ac.id

Abstrak

Evaluasi merupakan komponen penting dalam proses pembelajaran karena menentukan keberhasilan pengajaran. Penelitian ini bertujuan menganalisis kualita butir soal pada asesmen sumatif Pendidikan Agama Islam kelas VIII. Penelitian menggunakan pendekatan kuantitatif deskriptif untuk mengkaji kualitas instrumen tes yang meliputi validitas konstruk dan empiris, reliabilitas, tingkat kesukaran, daya beda, serta efektivitas pengecoh. Subek penelitian berjumlah 179 peserta didik kelas VIII. Data dikumpulkan melalui dokumentasi berupa soal dan lembar jawaban, kemudian dianalisis menggunakan IMB SPSS Statistik versi 27 dan ANATES V4. Hasil penelitian menunjukkan bahwa secara kognitif, soal didominasi level C3 (45%), C2 (30%), dan C4 (25%), tanpa adanya level C1, C5, dan C6. Secara empiris, seluruh butir soal 100% valid dengan nilai r_{tabel} 0,146. Reliabilitas sebesar 0,717 termasuk kategori reliabel. Tingkat kesukaran didominasi kategori sedang (65%) dan mudah (35%), tanpa adanya soal sukar. Daya beda tergolong baik hingga sangat baik (50%). Separuh pengecoh berfungsi efektif, sedangkan sisanya belum optimal. Dengan demikian, instrumen layak digunakan dengan perbaikan pada kualitas pengecoh dan pemerataan tingkat kognitif.

Kata kunci: Evaluasi Asesmen Sumatif, Kualitas Butir Soal, Pendidikan Agama Islam

Abstract

Evaluation is an essential component in the learning process, as it determines the effectiveness of instruction. This study aimed to analyze the quality of test items in the final semester of Islamic Religious Education for Grade VIII. A descriptive quantitative approach was employed to examine test quality, including construct and empirical validity, reliability, difficulty level, discrimination index, and distractor effectiveness. The subjects were 179 Grade VIII students.

Data were collected by documenting items and answer sheets and analyzed using IBM SPSS Statistics 27 and ANATES V4. The results indicated that cognitively, items were dominated by levels C3 (45%), C2 (35%), and C4 (25%), were no items at levels C1, C5, and C6. Empirically, all items were 100% valid, with a value of 0,146. The reliability coefficient was 0,717, indicating acceptable reliability. The difficulty level was mostly moderate (65%) and easy (35%), with no difficult items. The discrimination index ranged from good to very good (50% each). Half of the distractors functioned effectively, while the rest required improvement. Therefore, the test is feasible for use with revisions in distractor quality and cognitive level distribution., the test is feasible for use with revisions in distractor quality and cognitive level distribution.

Keywords: Assessment Sumative Evaluation, Item Quality, Islamic Religious Education

Pendahuluan

Pendidikan proses yang dirancang secara sadar dan sistematis untuk menciptakan pengalaman belajar yang bermakna, sehingga peserta didik mampu mengembangkan potensi secara optimal pada ranah kognitif, afektif, dan psikomotorik (Kemendikbudristek, 2022). Hal ini sejalan dengan Undang-Undang Sistem Pendidikan Nasional, yang menegaskan bahwa pendidikan tidak hanya berfungsi sebaagi sarana transfer ilmu pengetahuam, tetapi juga sebagai media pembentukan karakter dan kepribadian siswa. Melalui standarisasi pendidikan yang unggul, visi, untuk melahirkan sumber daya manusia yang mandiri, beretika, dan kontributif terhadap stabilitas nasional dapat terwujud. (BP et al., 2022). Oleh karena itu, pendidikan memiliki peran strategis dalam meningkatkan kualitas sumber daya manusia yang adaptif terhadap perkembangan zaman. Dalam konteks tersebut, Pendidikan Agama Islam menjadi mata pelajaran yang memiliki posisi penting dalam sistem pendidikan di Indonesia. Pembelajaran PAI tidak hanya terpaku pada penguasaan pengetahuan keagamaan secara teoritis, tetapi lebih mengutamakan proses nilai-nilai teologis dalam perilaku sosial yang mulia (Syafii et al., 2023). Namun, efektivitas internalisasi nilai tersebut sangat bergantung pada presisi mekanisme penilaian yang diterapkan. Kualitas evaluasi menjadi determinan utama dalam mendeteksi sejauh mana kompetensi spiritual dan etika tersebut diagnosis untuk memetakan keberhasilan target kurikuler sekaligus menjadi rujukan perbaikan pedagogis. Kualitas evaluasi menjadi determinan utama dalam mendeteksi sejauh mana kompetensi spiritual dan etika tersebut telah terserap secara autentik oleh siswa (Subakti, 2025).

Evaluasi pembelajaran marupakan komponen penting yang berfungsi untuk mengumpulkan informasi terkait

ketercapaian tujuan pembelajaran serta sebagai dasar dalam melakukan perbaikan proses pembelajaran (Indar Prabowo, 2024). Dalam praktiknya, evaluasi dapat dilakukan melalui teknik tes maupun non-tes, namun teknik tes masih menjadi instrumen utama dalam mengukur kemampuan kognitif peserta didik (Mukmin, 2024). Oleh karena itu, kualitas butir soal sebagai alat evaluasi menjadi sangat krusial. Sebuah instrumen evaluasi baru dianggap kredibel apabila memenuhi kriteria mencakup validitas, reliabilitas, indeks kesukaran, daya pembeda, serta fungsionalitas, sehingga mampu mengukur kemampuan peserta didik secara objektif dan akurat. (Santosa & Badawi, 2022).

Salah satu bentuk evaluasi adalah asesmen, yang dirancang untuk memantau progresivitas belajar sekaligus menyediakan umpan balik konstruktif bagi pendidik dan siswa (Suwandi, 2023). Secara umum, klasifikasi asesmen terbagi menjadi dua kategori: formatif, yang berintegrasi selama proses pembelajaran untuk perbaikan berkelanjutan; dan sumatif, yang bersifat konsklusif untuk memverifikasi pencapaian kompetensi di akhir periode (Firani Putri & Supratman Zakir, 2023). Fokus penyelidikan ini dikonsentrasikan pada ranah asesmen sumatif. Pada tingkat sekolah menengah pertama, penyusunan soal asesmen sumatif masih menghadapi kendala. Masih banyak tenaga pendidikan yang mengandalkan intuisi subjektif dalam merancang soal tanpa melalui prosedur analisis empiris yang terstandarisasi. Fenomena ini berdampak pada inkonsistensi mutu soal, seperti distribusi gradasi kesulitan yang tidak proporsional serta daya beda yang lemah. Jika kualitas teknis ini diabaikan, maka hasil penilaian yang diperoleh tidak akan mampu merepresentasikan kompetensi siswa yang sesungguhnya, sehingga memicu degradasi validitas penilaian secara sistemik (Setiawati et al., 2022).

Temuan awal di SMP Negeri 2 Kaligondang menunjukan bahwa meskipun penyusunan soal PAI dilakukan secara mandiri oleh guru, namun aktivitas tersebut belum disertai dengan evaluasi butir soal yang mendalam dan komprehensif. Hal ini menyebabkan kualitas instrumen evaluasi blum dapat diketahui secara pasti. Jika praktik ini terus berlanjut, data hasil belajar yang dihasilkan berisiko menimbulkan bias yang signifikan, yang pada akhirnya dapat menyesatkan pengambilan kebijakan edukasi dan penentuan kelulusan kompetensi siswa. Sejumlah penelitian terdahulu juga menunjukkan bahwa kualitas butir soal pada asesmen sumatif belum optimal terutama pada aspek validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh yang rendah. Realitas ini menegaskan bahwa analisis butir soal bukanlah sekadar prosedur administratif, melainkan langkah krusial untuk menjamin

AN-NAHDLOH Journal of Education and Islamic Studies

bahwa instrumen yang digunakan benar-benar memiliki daya ukur yang tepat. Tanpa analisis yang tajam, instrumen evaluasi hanya akan menjadi formalitas tanpa makna yang gagal memetakan kecakapan siswa secara transparan. Melalui pendekatan analisis holistik terhadap instrumen asesmen PAI dengan mensinergikan validitas konstruk dan empiris. Proses pengolahan data dilakukan secara presisi menggunakan bantuan teknologi perangkat lunak IMB SPSS *Statistics* 27 dan ANATES V4. Integrasi teknologi dalam analisis ini diproyeksikan mampu menyajikan data empiris yang lebih saintifik, mendalam, dan akurat mengenai kualitas teknis soal yang digunakan di sekolah tersebut. Secara spesifik, tujuan dari studi ini adalah untuk mengevaluasi secara tuntas kualitas butir soal asesmen sumatif pada mata pelajaran PAI untuk kelas VIII. Melalui identifikasi mendalam terhadap kelemahan dan kekuatan instrumen yang ada, penelitian ini diharapkan dapat memberikan kontribusi nyata bagi pengembangan standar evaluasi yang lebih reliabel dan profesional di masa depan

Metode Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan desain deskriptif yang bertujuan untuk menggambarkan kualitas butir soal secara objektif berdasarkan data empiris. Pendekatan kuantitatif dipilih karena penelitian ini berfokus pada pengolahan data numerik yang dianalisis secara statistik guna memperoleh kesimpulan yang akurat. Sementara itu, desain deskriptif digunakan untuk memaparkan karakteristik instrumen tes tanpa melakukan pengujian hipotesis. Penelitian dilaksanakan di SMP Negeri 2 Kaligondang dengan subjek penelitian sebanyak 179 peserta didik kelas VIII.

Data penelitian dikumpulkan melalui teknik dokumentasi yang meliputi soal asesmen sumatif, kunci jawaban, pedoman penskoran, serta lembar jawaban siswa. Selain itu, dilakukan wawancara terbatas dengan guru mata pelajaran untuk memperoleh informasi pendukung terkait proses penyusunan soal. Observasi non-sistematis juga dilakukan untuk melengkapi data mengenai pelaksanaan evaluasi pembelajaran. Analisis data dilakukan menggunakan statistik deskriptif dengan bantuan perangkat lunak IBM SPSS Statistics 27 dan ANATES V4. SPSS digunakan untuk menguji validitas empiris melalui teknik korelasi Product Moment serta menghitung reliabilitas menggunakan koefisien Alpha Cronbach, sedangkan ANATES digunakan untuk menganalisis karakteristik butir soal seperti tingkat kesukaran, daya beda, dan efektivitas pengecoh.

AN-NAHDLOH Journal of Education and Islamic Studies

Kriteria kualitas butir soal ditentukan berdasarkan parameter psikometrik yang meliputi validitas, reliabilitas, tingkat kesukaran, daya beda, dan efektivitas pengecoh. Validitas diukur dengan membandingkan nilai koefisien korelasi item dengan r tabel, reliabilitas ditentukan berdasarkan nilai koefisien Alpha, tingkat kesukaran diklasifikasikan menjadi mudah, sedang, dan sukar, daya beda dikategorikan dari buruk hingga sangat baik, serta efektivitas pengecoh dinilai berdasarkan kemampuan distraktor dalam menarik respons peserta didik. Melalui prosedur tersebut, penelitian ini menghasilkan gambaran empiris mengenai kualitas butir soal asesmen, yang dapat dijadikan dasar dalam perbaikan dan pengembangan instrumen evaluasi pembelajaran yang lebih berkualitas.

Hasil dan Pembahasan

(Analisis Validitas, Reliabilitas, Tingkat Kesukaran, Daya Beda, dan Efektivitas Pengecoh)

Evaluasi terhadap kualitas butir soal asesmen sumatif PAI kelas VIII di SMP Negeri 2 Kaligondang menunjukkan bahwa secara umum instrumen telah memenuhi kriteria psikometrik yang baik, meskipun masih terdapat beberapa aspek yang disempurnakan. Analisis dilakukan berdasarkan validitas konstruk dan empiris, reliabilitas, tingkat kesukaran, daya beda, serta efektivitas pengecoh.

1. Validitas (Konstruk dan Empiris)

a. Validitas Konstruk

Berdasarkan uji validitas konstruk, tingkat kognitif terhadap 20 butir soal Asesmen Sumatif mata pelajaran Pendidikan Agama Islam kelas VIII, distribusi soal didominasi oleh level kognitif menengah, yaitu C2 (memahami), C3 (mengaplikasikan), dan C4 (menganalisis). Secara kuantitatif, sebanyak 6 butir soal (30%) berada pada level C2 yang berfokus pada kemampuan memahami dan menginterpretasikan informasi, 9 butir soal (45%) pada level C3 yang menekankan kemampuan penerapan konsep dalam konteks baru, serta 5 butir soal (25%) pada level C4 yang mengukur kemampuan analisis melalui penguraian dan pengolahan informasi secara logis. Dominasi level C3 menunjukkan bahwa instrumen cenderung mengukur kemampuan berpikir tingkat menengah. Namun demikian, tidak ditemukannya soal pada level C1 (mengingat), C5 (mengevaluasi), dan C6 (mencipta) mengindikasikan bahwa cakupan pengukuran belum sepenuhnya mencerminkan keseluruhan spektrum Taksonomi Bloom, khususnya pada aspek *Higher Order Thinking Skills*

(HOTS). Oleh karena itu, meskipun distribusi soal tergolong cukup baik, diperlukan pengembangan instrumen dengan menambahkan variasi soal pada level kognitif rendah dan tinggi agar pengukuran kompetensi peserta didik menjadi lebih komprehensif.

b. Validitas Konstruk

Uji validitas empiris, instrumen Asesmen Sumatif mata pelajaran Pendidikan Agama Islam dilakukan menggunakan korelasi *Product Moment Pearson*. Penentuan validitas didasarkan pada perbandingan nilai r_{hitung}dengan r_{tabel}pada taraf signifikansi tertentu. Dengan jumlah responden sebanyak 179 siswa, diperoleh nilai $r_{tabel} = 0{,}146$.

Tabel 1. Hasil Uji Validitas

No. Soal	r hitung	r tabel
1	0,256	0,146
2	0,288	0,146
3	0,264	0,146
4	0,292	0,146
5	0,531	0,146
6	0,289	0,146
7	0,398	0,146
8	0,492	0,146
9	0,349	0,146
10	0,501	0,146
11	0,420	0,146
12	0,234	0,146
13	0,440	0,146
14	0,511	0,146
15	0,364	0,146
16	0,561	0,146
17	0,396	0,146
18	0,648	0,146
19	0,536	0,146
20	0,359	0,146

Hasil analisis menunjukkan bahwa seluruh butir soal memiliki nilai r_{hitung}yang lebih besar dari r_{tabel}, dengan rentang antara 0,234 hingga 0,648. Hal ini mengindikasikan bahwa setiap butir soal memiliki hubungan yang signifikan dengan skor total, sehingga memenuhi kriteria validitas empiris. Nilai korelasi tertinggi terdapat pada butir soal nomor 18, sedangkan nilai terendah pada nomor 12, namun keduanya tetap berada di atas ambang batas yang ditetapkan. Distribusi kelayakan butir soal menunjukkan bahwa seluruh item (100%) termasuk dalam kategori valid dan tidak memerlukan pengguguran. Temuan ini menegaskan bahwa instrumen yang digunakan telah memiliki tingkat kevalidan yang tinggi serta mampu merepresentasikan kompetensi yang diukur secara akurat.

Dengan demikian, instrumen Asesmen Sumatif dinyatakan layak digunakan sebagai alat evaluasi pembelajaran. Meskipun demikian, variasi nilai koefisien korelasi antarbutir tetap perlu diperhatikan sebagai dasar untuk penyempurnaan kualitas soal pada pengembangan instrumen berikutnya.

2. Reliabilitas

Uji reliabilitas instrumen dilakukan menggunakan koefisien Cronbach's Alpha untuk mengetahui tingkat konsistensi internal butir soal. Berdasarkan kriteria yang digunakan, suatu instrumen dinyatakan reliabel apabila memiliki nilai koefisien ≥ 0,70. Hasil analisis menunjukkan bahwa nilai koefisien reliabilitas instrumen sebesar 0,717.

Tabel 2. Hasil Uji Reliabilitas

Analisis	Kategori
Reliabilitas	R_1 = 0,717

Nilai tersebut berada di atas batas minimal yang ditetapkan, sehingga instrumen dapat dikategorikan memiliki reliabilitas tinggi. Dengan demikian, dapat disimpulkan bahwa seluruh butir soal dalam instrumen ASAS memiliki tingkat konsistensi yang baik dan layak digunakan sebagai alat ukur dalam menilai kemampuan peserta didik. Tingginya nilai reliabilitas ini menunjukkan bahwa instrumen mampu memberikan hasil yang stabil dan dapat dipercaya apabila digunakan dalam pengukuran yang sejenis.

3. Tingkat Kesukaran

Analisis tingkat kesukaran butir soal dilakukan untuk mengetahui proporsi kemampuan peserta didik dalam menjawab soal dengan benar. Indeks kesukaran diklasifikasikan menjadi tiga kategori, yaitu sukar, sedang, dan mudah, dengan prinsip bahwa semakin kecil nilai indeks, maka tingkat kesukaran soal semakin tinggi. Berdasarkan hasil analisis terhadap 20 butir soal, diperoleh rentang indeks kesukaran antara 0,31 hingga 0,93.

Tabel 3. Hasil Tingkat Kesukaran

No. Soal	B	JS	Indeks Kesukaran
1	92	179	0,51
2	109	179	0,61
3	73	179	0,41
4	146	179	0,82
5	121	179	0,68
6	112	179	0,63
7	144	179	0,80
8	158	179	0,88
9	107	179	0,60
10	166	179	0,93
11	155	179	0,87
12	55	179	0,31
13	113	179	0,63
14	97	179	0,54
15	83	179	0,46
16	110	179	0,61
17	111	179	0,62
18	144	179	0,80
19	132	179	0,74
20	120	179	0,67

Secara umum, soal cenderung berada pada kategori sedang hingga mudah. Tidak ditemukan butir soal dalam kategori sukar, yang menunjukkan bahwa sebagian besar soal relatif dapat dijangkau oleh peserta didik. Distribusi tingkat kesukaran menunjukkan bahwa sebanyak 13 butir soal (65%) termasuk

dalam kategori sedang, sedangkan 7 butir soal (35%) berada pada kategori mudah dan sangat mudah. Soal dengan tingkat kemudahan tertinggi terdapat pada nomor 8, 10, dan 11, sementara butir lainnya didominasi kategori sedang.

Dominasi soal kategori sedang menunjukkan bahwa instrumen telah cukup proporsional dalam mengukur kemampuan peserta didik, karena mampu membedakan tingkat penguasaan materi secara moderat. Namun demikian, tidak adanya soal kategori sukar serta masih ditemukannya sejumlah soal yang terlalu mudah mengindikasikan perlunya perbaikan pada beberapa butir soal agar distribusi tingkat kesukaran menjadi lebih seimbang. Dengan demikian, kualitas instrumen evaluasi dapat ditingkatkan untuk mengukur kemampuan peserta didik secara lebih komprehensif.

4. Daya Beda

Analisis daya beda dilakukan untuk mengukur kemampuan butir soal dalam membedakan peserta didik yang berkemampuan tinggi dan rendah. Indeks daya beda (D) dihitung berdasarkan selisih proporsi jawaban benar antara kelompok atas dan kelompok bawah. Kriteria interpretasi meliputi kategori sangat jelek (D < 0,00), jelek (0,00–0,20), cukup (0,20–0,40), baik (0,40–0,70), dan sangat baik (0,70–1,00). Hasil analisis menunjukkan bahwa nilai indeks daya beda berkisar antara 0,21 hingga 0,67.

Tabel 4. Hasil Daya Beda

No. Soal	BA	BB	JA	JB	D
1	36	19	48	48	0,35
2	34	22	48	48	0,25
3	28	15	48	48	0,27
4	42	32	48	48	0,21
5	44	14	48	48	0,63
6	40	22	48	48	0,38
7	46	29	48	48	0,35
8	47	32	48	48	0,31
9	36	16	48	48	0,42
10	48	37	48	48	0,23
11	48	33	48	48	0,31
12	27	10	48	48	0,35
13	40	18	48	48	0,46
14	42	12	48	48	0,63
15	33	12	48	48	0,44
16	46	14	48		0,67
17	42	17	48		0,52
18	48	20	48		0,58
19	47	16	48		0,65
20	40	19	48		0,44

Secara umum, seluruh butir soal memiliki nilai diskriminasi positif, yang berarti mampu membedakan kemampuan peserta didik secara efektif. Tidak ditemukan soal dengan kategori jelek maupun sangat jelek. Distribusi daya beda menunjukkan bahwa

AN-NAHDLOH Journal of Education and Islamic Studies

sebanyak 4 butir soal (20%) berada pada kategori cukup, yaitu nomor 2, 3, 4, dan 10. Selanjutnya, 6 butir soal (30%) termasuk kategori baik, yaitu nomor 1, 6, 7, 8, 11, dan 12. Adapun 10 butir soal (50%) lainnya berada pada kategori sangat baik, yaitu nomor 5, 9, 13, 14, 15, 16, 17, 18, 19, dan 20.

Dominasi kategori sangat baik menunjukkan bahwa sebagian besar soal memiliki kemampuan tinggi dalam membedakan tingkat kompetensi peserta didik. Secara keseluruhan, kualitas daya beda instrumen dapat dikatakan baik dan layak digunakan sebagai alat evaluasi. Meskipun demikian, beberapa butir soal dengan kategori cukup masih perlu diperbaiki agar kemampuan diskriminatifnya dapat ditingkatkan, sehingga instrumen menjadi lebih optimal dalam mengukur perbedaan kemampuan peserta didik secara lebih akurat.

5. Efektivitas Pengecoh

Analisis efektivitas pengecoh dilakukan untuk mengetahui sejauh mana opsi jawaban selain kunci (distraktor) mampu berfungsi dengan baik dalam menarik perhatian peserta didik. Instrumen menggunakan empat pilihan jawaban dengan tiga pengecoh pada setiap butir soal. Pengecoh dikatakan efektif apabila dipilih oleh peserta didik dalam proporsi yang memadai, sedangkan pengecoh yang jarang atau tidak dipilih menunjukkan ketidakefektifan. Hasil analisis menunjukkan bahwa tidak seluruh distraktor berfungsi secara optimal. Dari 20 butir soal yang dianalisis, sebanyak 10 butir (50%) memiliki pengecoh efektif, yaitu pada nomor 1, 6, 12, 13, 14, 15, 16, 17, 19, dan 20. Sementara itu, 10 butir lainnya (50%) termasuk dalam kategori pengecoh tidak efektif, yaitu nomor 2, 3, 4, 5, 7, 8, 9, 10, 11, dan 18.

Tabel 5. Hasil Efektivitas Pengecoh

No	IP A	IP B	IP C	IP D
1	13,4%	30,2%	*	5,0%
2	26,3%	*	8,9%	3,9%
3	*	8,4%	49,7%	1,1%
4	*	1,1%	11,2%	6,1%
5	*	24,0%	6,1%	2,2%
6	12,8%	*	19,0%	5,6%
7	10,1%	*	6,7%	2,8%
8	1,7%	2,2%	*	7,8%
9	33,0%	2,2%	*	7,8%
10	1,1%	3,4%	2,8%	*
11	4,5%	*	6,1%	2,8%
12	14,5%	11,2%	*	43,6%
13	10,6%	*	11,2%	15,1%
14	*	12,3%	14,0%	19,6%
15	9,5%	16,2%	*	27,9%
16	*	20,7%	11,2%	6,7%
17	12,8%	14,0%	11,2%	*
18	2,2%	6,1%	11,2%	*
19	7,3%	10,1%	*	8,9
20	11,2%	11,2%	10,6%	*

Ketidakefektifan pengecoh ditunjukkan oleh rendahnya persentase pemilihan atau bahkan tidak adanya peserta didik yang memilih opsi tersebut. Hal ini

AN-NAHDLOH Journal of Education and Islamic Studies

mengindikasikan bahwa beberapa distraktor kurang mampu mengecoh peserta didik atau terlalu mudah dieliminasi. Sebaliknya, pengecoh yang efektif menunjukkan bahwa opsi jawaban mampu berfungsi sebagai alternatif yang masuk akal bagi peserta didik yang belum menguasai materi. Secara keseluruhan, efektivitas pengecoh dalam instrumen ini masih berada pada kategori sedang karena proporsi antara pengecoh efektif dan tidak efektif seimbang (50% : 50%). Oleh karena itu, diperlukan perbaikan pada butir soal yang memiliki pengecoh tidak efektif, khususnya dengan menyusun opsi jawaban yang lebih variatif, logis, dan sesuai dengan karakteristik kesalahan umum peserta didik, agar kualitas instrumen evaluasi menjadi lebih optimal. Hasil penelitian menunjukkan bahwa instrumen asesmen sumatif telah memenuhi sebagian besar kriteria kualitas butir soal, terutama pada aspek validitas, reliabilitas, daya beda. Namun demikian, masih diperlukan perbaikan soal yang mencakup kemampuan berpikir tingkat tinggi. Dengan demikian, instrumen evaluasi yang dihasilkan diharapkan mampu memberikan pengukuran yang lebih komprehensif dan akurat terhadap kompetensi siswa.

Kesimpulan

Penelitian ini bertujuan untuk menganalisis kualitas butir soal asesmen sumatif mata pelajaran Pendidikan Agama Islam kelas VIII di SMP Negeri 2 Kaligondang berdasarkan aspek validitas, reliabilitas, tingkat kesukaran, daya pembedan, dan efektivitas pengecoh. Hasil penelitian menunjukkan bahwa secara umum instrumen yang digunakan telah memnuhi kriteria kualitas yang baik. Seluruh butir soal dinyatakan valid secara empiris dan konstruk, serta memiliki konsistensi internal yang tinggi sebagaimana ditunjukkan oleh nilai reliabilitas yang melampaui batas minimal. Dari segi tingkat kesukaran, distribusi soal tergolong cukup seimbang dengan dominasi kategori sedang, meskipun belum terdapat butir soal dengan tingkat kesukaran tinggi. Daya pembeda instrumen juga menunjukkan kategori baik hingga sangat baik, sehingga mampu membedakan kemampuan peserta didik secara efektif. Namun demikian, efektivitas pengecoh masih perlu ditingkatkan karena sebagian pilihan jawaban belum berfungsi secara optimal.

Secara keseluruhan, instrumen asesmen PAI kelas VIII di SMP Negeri 2 Kaligondang telah memenuhi standar kualitas yang cukup baik, namun masih memerlukan perbaikan terutama pada penyusunan pengecoh dan pemerataan tingkat kognitif soal. Oleh karena itu, penelitian selanjutnya disarankan untuk mengembangkan instrumen evaluasi yang lebih bervariatif dengan memperhatikab distribusi tingkat kesukaran dan kualitas

AN-NAHDLOH
Journal of Education and Islamic Studies

distraktor, serta memanfaatkan analisis berbasis teknologi secara berkelanjutan guna meningkatkan akurasi penilaian pembelajaran.

Referensi

BP, A. R., Munandar, S. A., Fitriani, A., Karlina, Y., & Yumriana. (2022). Pengertian pendidikan, ilmu pendidikan dan unsur-unsur pendidikan. *Al Urwatul Wutsqa: Kajian Pendidikan Islam*, *2*(1), 1–8.

Firani Putri, & Supratman Zakir. (2023). Mengukur Keberhasilan Evaluasi Pembelajaran: Telaah Evaluasi Formatif Dan Sumatif Dalam Kurikulum Merdeka. *Dewantara : Jurnal Pendidikan Sosial Humaniora*, *2*(4), 172–180. https://doi.org/10.30640/dewantara.v2i4.1783

Indar Prabowo. (2024). Analisis Kualitas Soal Asesmen Sumatif Pendidikan Agama Islam dan Budi Pekerti. *Jurnal Ilmiah Insan Mulia*, *1*(1), 1–7. https://doi.org/10.59923/jiim.v1i1.155

Kemendikbudristek. (2022). Peraturan Menteri Pendidikan, Kebudayaan, Riset, Dan Teknologi No 21 Tahun 2022 Tentang Standar Penilaian Pendidikan Pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, Dan Jenjang Pendidikan Menengah. *Kementerian Pendidikan Dan Kebudayaan Riset Dan Teknologi, 9*.

Mukmin, N. (2024). Integrasi Penilaian Tes dan Non-Tes dalam Pendidikan Agama Islam : Menuju Evaluasi Holistik untuk Pembelajaran Berkelanjutan. *Journal of Instructional and Development Researches*, *4*(5), 370–379. https://www.journal.iel-education.org/index.php/JIDeR

Santosa, S., & Badawi, J. A. (2022). Analisis Butir Soal Pilihan Ganda Tema Pertumbuhan Dan Perkembangan Makhluk Hidup Kelas III Madrasah Ibtidaiyah. *Jurnal Basicedu*, *6*(2), 1678–1686. https://doi.org/https://doi.org/10.31004/basicedu.v6i2.2206

Setiawati, M. N., Purwanto, & Ramalis, T. R. (2022). Analisis Parameter Tes Penilaian Akhir Semester Fisika Kelas X dengan Teori Respon Butir. *Wahana Pendidikan Fisika*, *7*(1), 1–10.

Subakti, widya. (2025). Analisis Kualitas Butir Soal Pada Uji Coba Evaluasi Pembelajaran Matematika Tingkat Sma. *CENDEKIA : Jurnal Ilmu Pengetahuan*, *5*(1). https://jurnalp4i.com/index.php/cendekia

Suwandi, I. N. (2023). The Roles of Teachers and Shcools in Conducting Effective Classrom Assesment. *International Journal of Social Science*, *3*(1), 45–50.

Syafii, A., Bahar, Shobicah, & Muharam, A. (2023). Pengukuran Indeks Mutu Pendidikan Berbasis Standar Nasional. *Jurnal Multidisiplin Indonesia*, *2*(7), 1697–1701.

AN-NAHDLOH
Journal of Education and Islamic Studies