Afeksi: Jurnal Penelitian dan Evaluasi Pendidikan
Volume 6 Nomor 6 Tahun 2025
https://afeksi.id/jurnal/index.php/afeksi
e-ISSN: 2745-9985

Analisis Butir Soal Berbasis HOTS pada Mata Pelajaran Al-Qur'an Hadis dengan Menggunakan Program Anates

Rais Kamil Taqiy[1], Heny Narendrany Hidayati[2]
UIN Syarif Hidayatullah Jakarta, Indonesia
Email : Raiskamiltaqiy1611@gmail.com, heny.fitk@yahoo.com

Abstrak
Penelitian ini bertujuan untuk menganalisis kualitas butir soal berbasis *HOTS* pada mata pelajaran Al-Qur'an Hadis kelas X. Analisis butir soal dilakukan melalui uji coba instrumen dengan menggunakan Anates. Penelitian menggunakan pendekatan kuantitatif deskriptif. Subjek penelitian 30 peserta didik kelas X MA Al-Fathiyah Jakarta Timur yang mengerjakan tes soal pilihan ganda sebanyak 30 butir soal. Data dianalisis untuk mengetahui tingkat kesukaran, daya pembeda, fungsi distraktor, validitas butir soal, serta validitas dan reliabilitas instrumen tes. Hasil analisis menunjukkan bahwa sebagian besar butir soal berada pada kategori tingkat kesukaran sedang hingga sukar. Ditinjau dari daya pembeda, mayoritas butir soal memiliki kemampuan daya beda yang baik . Selain itu, hasil analisis fungsi distraktor memperlihatkan sebagian besar pilihan pengecoh dapat berfungsi dengan baik. Hasil uji validitas butir soal menunjukkan mayoritas butir soal berada pada kategori valid dengan tingkat signifikan hingga sangat signifikan. Selanjutnya hasil uji validitas instrumen menunjukkan nilai koefisien sebesar 0,94 yang termasuk dalam kategori sangat tinggi. Hasil uji reliabilitas instrumen juga menunjukkan kategori sangat tinggi yaitu sebesar 0,97. Berdasarkan hasil analisis butir soal tersebut, dapat disimpulkan bahwa instrumen tes mata pelajaran Al-Qur'an Hadits berbasis *Higher Order Thinking Skills* (HOTS) memiliki kualitas yang baik dan layak digunakan sebagai alat evaluasi pembelajaran.
Kata Kunci: *Al-Qur'an Hadis, analisis butir soal, HOTS, ANATES*

PENDAHULUAN

Evaluasi pembelajaran merupakan komponen esensial dalam proses pendidikan karena berfungsi sebagai alat untuk menilai ketercapaian tujuan pembelajaran serta memberikan umpan balik terhadap efektivitas strategi pembelajaran yang diterapkan. Dalam konteks pendidikan agama Islam, khususnya pada mata pelajaran Al-Qur'an Hadis di Madrasah Aliyah, evaluasi tidak hanya diarahkan pada penguasaan pengetahuan faktual, tetapi juga pada kemampuan berpikir tingkat tinggi peserta didik dalam memahami, menafsirkan, dan mengaitkan nilai-nilai keislaman dengan konteks kehidupan nyata. Oleh karena itu, kualitas instrumen evaluasi menjadi faktor penting dalam menghasilkan penilaian yang objektif dan bermakna.

Evaluasi pembelajaran tidak hanya berfungsi untuk mengetahui pencapaian hasil belajar peserta didik, tetapi juga menjadi dasar dalam pengambilan keputusan dan perbaikan proses pembelajaran secara berkelanjutan. Oleh karena itu, kualitas instrumen evaluasi menjadi aspek penting yang menentukan akurasi dan kebermaknaan hasil penilaian (Hidayati, 2024**;** Hidayati, 2025). Berbagai penelitian dalam bidang evaluasi pendidikan menunjukkan bahwa instrumen tes yang digunakan di satuan pendidikan masih

menghadapi persoalan mendasar, terutama terkait kualitas butir soal. Hasil penelitian pada jurnal pendidikan mengungkapkan bahwa banyak instrumen evaluasi belum melalui analisis empiris secara memadai, sehingga menghasilkan butir soal dengan tingkat kesukaran yang kurang proporsional, daya pembeda yang rendah, serta distraktor yang tidak berfungsi secara optimal (Saputra et al., 2022). Kondisi tersebut berdampak pada ketidakakuratan hasil penilaian dan berpotensi tidak merepresentasikan kemampuan peserta didik secara objektif.

Permasalahan kualitas butir soal juga banyak ditemukan pada tes pilihan ganda yang digunakan dalam evaluasi pembelajaran. Penelitian analisis butir soal menunjukkan bahwa instrumen evaluasi yang tidak diuji secara empiris cenderung didominasi oleh soal dengan level kognitif rendah, sehingga kurang mampu mengukur kemampuan berpikir tingkat tinggi peserta didik. Padahal, tuntutan pembelajaran abad ke-21 menekankan pentingnya pengembangan *Higher Order Thinking Skills* (HOTS), yang mencakup kemampuan analisis, evaluasi, dan penalaran kritis. Oleh karena itu, instrumen evaluasi yang berkualitas perlu disusun dan diuji secara sistematis agar selaras dengan tujuan pembelajaran berbasis HOTS.

Dalam konteks pendidikan agama Islam, khususnya mata pelajaran Al-Qur'an Hadis, analisis kualitas butir soal memiliki urgensi yang lebih besar karena materi pembelajaran tidak hanya bersifat konseptual, tetapi juga normatif dan aplikatif. Instrumen evaluasi yang kurang berkualitas berpotensi gagal mengukur pemahaman peserta didik terhadap kandungan nilai Al- Qur'an dan Hadis secara komprehensif. Penelitian-penelitian terdahulu menegaskan bahwa analisis butir soal, yang mencakup tingkat kesukaran, daya pembeda, fungsi distraktor, validitas, dan reliabilitas, merupakan langkah strategis dalam menjamin mutu instrumen evaluasi dan meningkatkan akurasi penilaian hasil belajar (Saputra et al., 2022)

Salah satu pendekatan yang banyak digunakan dalam analisis empiris butir soal adalah pemanfaatan perangkat lunak analisis tes, seperti program ANATES. Penggunaan ANATES memungkinkan peneliti memperoleh gambaran kuantitatif yang objektif mengenai karakteristik setiap butir soal, sehingga memudahkan pendidik dalam melakukan evaluasi dan perbaikan instrumen secara berbasis data. Meskipun demikian, kajian yang secara khusus menganalisis kualitas butir soal mata pelajaran Al-Qur'an Hadis di tingkat Madrasah Aliyah dengan pendekatan HOTS dan berbantuan program ANATES masih relatif terbatas.

Sedangkan evaluasi pendidikan merupakan proses sistematis yang melibatkan pengumpulan, analisis, dan penafsiran data untuk mengetahui ketercapaian tujuan pembelajaran serta sebagai dasar pengambilan keputusan pendidikan. Evaluasi yang dirancang dengan baik memungkinkan pendidik memperoleh gambaran yang lebih komprehensif mengenai efektivitas pembelajaran dan kualitas instrumen yang digunakan. Oleh karena itu, instrumen evaluasi memiliki peran strategis dalam memastikan bahwa data yang diperoleh benar-benar merepresentasikan kemampuan peserta didik secara akurat (Kamal & Nursikin, 2025).

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan untuk menganalisis kualitas empiris butir soal mata pelajaran Al-Qur'an Hadis berbasis *Higher Order Thinking*

Skills (HOTS) dengan menggunakan program ANATES. Analisis difokuskan pada lima aspek utama, yaitu tingkat kesukaran, daya pembeda, fungsi distraktor, validitas, dan reliabilitas instrumen. Hasil penelitian ini diharapkan dapat memberikan kontribusi empiris dalam pengembangan instrumen evaluasi pembelajaran Al-Qur'an Hadis yang lebih berkualitas serta menjadi rujukan bagi pendidik dalam menyusun evaluasi pembelajaran yang mampu mengukur kemampuan berpikir tingkat tinggi peserta didik secara lebih akurat.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan jenis penelitian deskriptif. Pendekatan kuantitatif dipilih karena penelitian ini berfokus pada pengolahan dan analisis data numerik untuk menggambarkan kualitas empiris instrumen tes berdasarkan hasil uji coba. Penelitian deskriptif bertujuan untuk memberikan gambaran objektif mengenai karakteristik instrumen tanpa memberikan perlakuan tertentu terhadap subjek penelitian (Sugiyono, 2019).

Subjek penelitian ini adalah peserta didik kelas X MA Al-Fathiyah yang mengikuti kegiatan uji coba instrumen dengan jumlah responden sebanyak 30 peserta didik. Teknik pengambilan sampel yang digunakan adalah total sampling, yaitu seluruh anggota populasi dijadikan sampel penelitian karena jumlah populasi relatif terbatas dan memungkinkan untuk dijangkau secara keseluruhan (Sugiyono, 2019).

Penelitian dilaksanakan melalui beberapa tahapan, meliputi penyusunan instrumen tes, pelaksanaan uji coba instrumen, serta pengolahan dan analisis data hasil uji coba. Tahapan penelitian ini disusun secara sistematis untuk memperoleh gambaran kualitas butir soal secara empiris sesuai dengan tujuan penelitian.

Instrumen penelitian berupa tes objektif dalam bentuk pilihan ganda yang berjumlah 30 butir soal mata pelajaran Al-Qur'an Hadis. Instrumen disusun berdasarkan kisi-kisi yang mengacu pada tujuan pembelajaran, capaian pembelajaran, serta materi yang telah diajarkan. Evaluasi kualitas instrumen difokuskan pada lima aspek utama, yaitu tingkat kesukaran, daya pembeda, fungsi distraktor, validitas, dan reliabilitas tes. Aspek-aspek tersebut merupakan indikator utama dalam analisis kualitas butir soal sebagaimana digunakan dalam berbagai penelitian evaluasi pembelajaran (Muniroh, 2024; Rahman & Khalidi, 2025).

Validitas instrumen dalam penelitian ini ditentukan melalui uji validitas empiris dengan mengkorelasikan skor setiap butir soal terhadap skor total tes. Butir soal yang memiliki koefisien korelasi signifikan dinyatakan valid secara empiris dan layak digunakan sebagai alat evaluasi hasil belajar peserta didik (Rahman & Khalidi, 2025). Penelitian ini memfokuskan validitas instrumen pada validitas empiris berdasarkan data hasil uji coba, sedangkan validitas isi dikonstruksi melalui penyusunan kisi-kisi yang mengacu pada indikator pembelajaran dan materi ajar.

Pengumpulan data dilakukan melalui hasil respons peserta didik terhadap tes yang dikerjakan secara tatap muka melalui *Google Form*. Data yang diperoleh berupa skor jawaban peserta didik pada setiap butir soal, yang selanjutnya digunakan sebagai dasar dalam proses analisis kualitas instrumen tes. Penggunaan data respons peserta didik memungkinkan

peneliti memperoleh informasi empiris mengenai kinerja setiap butir soal secara objektif dan terukur.

Analisis data dilakukan secara kuantitatif dengan bantuan program ANATES untuk memperoleh informasi mengenai tingkat kesukaran, daya pembeda, fungsi distraktor, validitas, dan reliabilitas butir soal. Hasil analisis tersebut kemudian ditafsirkan secara deskriptif untuk menggambarkan kualitas instrumen tes yang digunakan.

HASIL DAN PEMBAHASAN

Instrumen tes Al-Qur'an Hadis disusun berbasis *Higher Order Thinking Skills* (HOTS), berada pada level penalaran yaitu C4, C5 dan C6, dengan pokok bahasan Al Qur'an adalah Wahyu Allah SWT dan Hadits sumber ajaran islam. Untuk mengetahui kualitas butir soal maka dilakukan analisis butir soal yang meliputi taraf sukar, daya beda, fungsi distraktor dan validitas butir, serta validitas dan reliabilitas instrumen.

Analisis kualitas instrumen tes melalui pengujian validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan fungsi distraktor merupakan bagian penting dalam evaluasi pembelajaran karena menentukan kelayakan instrumen sebagai alat ukur hasil belajar peserta didik. Instrumen yang dianalisis secara empiris memungkinkan pendidik memperoleh gambaran objektif mengenai kualitas butir soal yang digunakan sebagai dasar evaluasi dan perbaikan pembelajaran (Hidayati, 2024).

Bagian ini menyajikan hasil analisis kualitas empiris butir soal mata pelajaran Al-Qur'an Hadis berdasarkan data hasil uji coba instrumen menggunakan program ANATES. Instrumen yang dianalisis terdiri atas 30 butir soal pilihan ganda yang diberikan kepada 30 peserta didik. Analisis butir soal difokuskan pada enam aspek utama, yaitu tingkat kesukaran, daya pembeda, fungsi distraktor, validitas butir soal, serta validitas dan reliabilitas instrumen tes. Hasil analisis butir soal menunjukkan bahwa instrumen tes matapelajaran Al-Qur'an Hadis berkualitas.

Dapat dibuktikan berdasarkan analisis taraf sukar butir soal yang menunjukan mayoritas butir pada kategori sedang dan sukar. Daya pembeda mayoritas butir soal memiliki kemampuan yang baik dalam membedakan peserta didik berdasarkan tingkat penguasaan kompetensinya. Hasil analisis fungsi distraktor juga memperlihatkan bahwa sebagian besar pilihan pengecoh telah berfungsi secara efektif. Hasil uji validitas butir soal menunjukkan bahwa mayoritas butir soal berada pada kategori valid dengan tingkat signifikan hingga sangat signifikan. Sementara itu, hasil uji validitas instrumen menunjukkan nilai koefisien sebesar 0,94 yang termasuk dalam kategori sangat tinggi. Hasil uji reliabilitas instrumen juga menunjukkan koefisien reliabilitas sebesar 0,97 yang tergolong sangat tinggi.

Tingkat Kesukaran Butir Soal

Berdasarkan hasil pengolahan data menggunakan program ANATES, diketahui bahwa tingkat kesukaran butir soalpada mata pelajaran Al-Qur'an Hadis kelas X berada pada kategori sedang dan sukar. Dari total 30 butir soal yang dianalisis, mayoritas tergolong dalam kategori sedang, sedangkan

sebagian lainnya termasuk kategori sukar. Tidak terdapat butir soal yang masuk dalam kategori mudah maupun sangat mudah. Hasil hitung menunjukan mayoritas soal berada pada kategori sedang dan sukar.

Kriteria tingkat kesukaran suatu item soal dibuat klasifiksasi, yaitu: Indeks kesukaran 0,00-0,30 tergolong sukar, Indeks kesukaran 0,31-0,70 tergolong sedang, dan Indeks kesukaran 0,71-1,00 tergolong mudah Kadir, A. (2015). Ringkasan hasil analisis disajikan pada Tabel 1.

Tabel 1. Tingkat kesukaran Butir Soal

Kategori	Rentang (%)	Jumlah Butir	Peesentase	Nomor Butir
Mudah	71-100	0	0,00 %	-
Sedang	31-70	28	93,33 %	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 28, 29, 30
Sukar	0-30	2	6,67 %	21, 27

Berdasarkan Tabel 1, butir soal didominasi kategori sedang (93,33%), sedangkan butir kategori sukar sebanyak 6,67% (butir 21 dan 27). Tidak terdapat butir pada kategori mudah maupun sangat mudah. Komposisi ini menunjukkan tingkat kesulitan instrumen cenderung proporsional untuk mengukur kemampuan peserta didik. Distribusi tingkat kesukaran butir soal yang didominasi oleh kategori sedang dan sukar menunjukkan bahwa instrumen tes mata pelajaran Al-Qur'an Hadis ditujukan untuk mengukur kemampuan berpikir tingkat tinggi peserta didik. Beberapa penelitian menunjukkan bahwa soal yang dirancang untuk mengukur kemampuan berpikir tingkat tinggi (HOTS) cenderung menuntut proses kognitif kompleks seperti analisis, evaluasi, dan pemecahan masalah, sehingga soal ber-tingkat kesukaran sedang hingga sukar lebih efektif dalam mengungkap kompetensi peserta didik pada level berpikir tinggi dibanding soal yang hanya menuntut kemampuan mengingat atau memahami saja (Wikipedia, 2024; Purwati et al., 2021). Instrumen evaluasi dengan karakteristik tersebut dinilai lebih selaras dengan prinsip penilaian berbasis *Higher Order Thinking Skills* (HOTS), yang menuntut peserta didik untuk tidak sekadar mengingat informasi, tetapi juga mengolah, mengaitkan, dan menerapkan pengetahuan dalam konteks yang lebih kompleks (Muniroh, 2024; Rahman & Khalidi, 2025).

Daya Pembeda Butir Soal

Indeks daya pembeda digunakan untuk mengetahui kemampuan suatu butir soal dalam membedakan peserta didik yang memiliki kemampuan tinggi dan rendah. Daya pembeda dinyatakan dalam bentuk indeks (D) dengan rentang nilai −1,00 hingga +1,00. Semakin besar nilai D, semakin baik kemampuan butir soal dalam membedakan tingkat kemampuan peserta didik.

Kriteria indeks daya pembeda umumnya diklasifikasikan sebagai berikut: nilai $D \leq 0{,}00$ termasuk kategori sangat jelek; $0{,}00 < D \leq 0{,}20$ tergolong jelek; $0{,}20 < D \leq 0{,}40$ termasuk kategori cukup; 0,40

$< D \leq 0{,}70$ tergolong baik; dan $D > 0{,}70$ termasuk kategori sangat baik. Butir soal dengan indeks daya pembeda rendah atau bernilai negatif dinyatakan kurang layak digunakan

karena tidak mampu membedakan kemampuan peserta didik secara efektif (Mania et al., 2020). Ringkasan daya beda disajikan pada Tabel 2.

Tabel 2. Daya Beda Butir Soal

Kategori	Indeks DP	Jumlah Butir	Presentase	Nomor Butir
Sangat Baik	> 0,70	17	56,67 %	3, 4, 7, 8, 9, 10, 12, 15, 16, 18, 19, 20, 21, 23, 26, 28, 30
Baik	0,40 - 0,70	11	36,67 %	1, 2, 5, 11, 13, 14, 17, 22, 24, 27, 29
Cukup	0,20 - 0,40	2	6,67 %	6, 25
Buruk	0,00 - 0,20	0	0,00 %	-
Sangat Buruk	< 0,00	0	0,00 %	-

Hasil analisis menunjukkan 17 butir (56,67%) berada pada kategori sangat baik dan 11 butir (36,67%) berada pada kategori baik. Terdapat 2 butir (6,67%), yaitu butir 6 dan 25, berada pada kategori cukup sehingga perlu ditinjau kembali stimulus atau alternatif jawabannya agar kemampuan diskriminatif meningkat.

Daya pembeda merupakan indikator penting dalam menentukan kualitas butir soal karena menunjukkan kemampuan instrumen evaluasi dalam membedakan tingkat penguasaan materi peserta didik. Butir soal dengan daya pembeda rendah menandakan bahwa instrumen belum bekerja secara optimal, sehingga perlu dilakukan revisi agar penilaian hasil belajar dapat memberikan informasi yang lebih akurat mengenai perbedaan kemampuan peserta didik (Rahman & Khalidi, 2025).

Fungsi Distraktor

Pada tes objektif berbentuk pilihan ganda, alternatif jawaban umumnya terdiri atas satu kunci jawaban dan beberapa opsi pengecoh (distraktor). Distraktor berfungsi untuk mengecoh peserta didik yang belum menguasai materi, sehingga kualitas distraktor menjadi salah satu indikator penting dalam menilai efektivitas butir soal. Penelitian dalam bidang evaluasi pendidikan menunjukkan bahwa distraktor yang baik adalah distraktor yang dipilih oleh peserta didik dengan kemampuan rendah dan relatif tidak dipilih oleh peserta didik dengan kemampuan tinggi, sehingga mampu membedakan tingkat penguasaan materi peserta didik secara lebih akurat (Rahman & Khalidi, 2025).

Hasil analisis kualitas distraktor pada instrumen tes HOTS menunjukkan bahwa secara umum pengecoh telah berfungsi dengan baik. Hal ini ditunjukkan oleh dominasi butir soal yang memiliki distraktor pada kategori baik (+) dan sangat baik (++). Kondisi tersebut mengindikasikan bahwa alternatif jawaban yang disusun bersifat relatif homogen, logis, dan tidak mudah ditebak, sehingga mampu mengecoh peserta didik yang belum menguasai materi secara optimal. Temuan ini sejalan dengan penelitian sebelumnya yang menegaskan bahwa distraktor yang berfungsi dengan baik berkontribusi terhadap meningkatnya kualitas instrumen evaluasi berbasis HOTS (Muniroh, 2024; Rahman & Khalidi, 2025).

Beberapa penelitian merekomendasikan bahwa perbaikan distraktor dapat dilakukan

dengan menyusun alternatif jawaban yang lebih sejajar, logis, serta relevan dengan miskonsepsi yang sering dialami peserta didik. Dengan demikian, distraktor tidak hanya berfungsi sebagai pelengkap pilihan jawaban, tetapi juga sebagai indikator kemampuan peserta didik dalam memahami konsep yang diujikan.

Berdasarkan hasil uji kualitas pengecoh menggunakan program ANATES terhadap instrumen tes HOTS, fungsi distraktor dapat diklasifikasikan ke dalam beberapa kriteria, yaitu sangat baik, baik, buruk, dan sangat buruk. Pengelompokan ini didasarkan pada kualitas distraktor terendah yang terdapat pada setiap butir soal, sebagaimana disajikan pada Tabel 3.

Tabel 3. Hasil Uji Fungsi Distraktor Instrumen Tes HOTS

Kriteria Fungsi Distraktor	Nomor Butir Soal
Sangat Baik ++	2, 3, 4, 5, 6, 8, 9, 12, 14, 16, 18, 19, 20, 21, 22, 24, 26, 28, 29
Baik +	1, 7, 10, 11, 15, 17, 23, 25, 27, 30
Buruk -	13
Sangat Buruk -	–

Berdasarkan Tabel 3, sebagian besar butir soal memiliki fungsi distraktor pada kategori sangat baik dan baik. Hal ini menunjukkan bahwa pilihan jawaban yang disediakan telah mampu mengecoh peserta didik yang belum menguasai materi serta tidak menimbulkan pola jawaban yang terlalu mencolok. Dengan demikian, secara umum distraktor pada instrumen tes HOTS telah berfungsi sesuai dengan prinsip evaluasi pembelajaran yang baik.

Hanya terdapat satu butir soal yang memiliki distraktor dengan kategori buruk, yaitu butir soal nomor 13. Distraktor pada butir tersebut menunjukkan kualitas yang rendah karena terdapat alternatif jawaban yang tidak dipilih oleh peserta didik atau terlalu jelas kesalahannya, sehingga mengurangi efektivitas butir soal. Kondisi ini berpotensi menurunkan kualitas instrumen apabila tidak dilakukan perbaikan. Oleh karena itu, butir soal dengan fungsi distraktor yang kurang optimal perlu direvisi melalui perbaikan redaksi dan homogenitas pilihan jawaban agar kualitas instrumen tes dapat ditingkatkan secara maksimal (Rahman & Khalidi, 2025).

Validitas Butir Soal

Validitas dan reliabilitas instrumen merupakan fondasi utama dalam penelitian kuantitatif karena menentukan akurasi serta kredibilitas data yang dihasilkan. Instrumen yang tidak melalui pengujian validitas dan reliabilitas berpotensi menghasilkan data yang bias dan mengganggu ketepatan interpretasi hasil penelitian.Validitas dan reliabilitas instrumen merupakan dua aspek utama yang menentukan kualitas pengukuran dalam evaluasi pembelajaran. Instrumen yang valid dan reliabel tidak hanya mampu mengukur kompetensi secara tepat, tetapi juga memberikan dasar yang kuat bagipengambilan keputusan dalam pembelajaran (Hidayati, 2024). Validitas butir soal berkaitan dengan sejauh mana suatu butir mampu mengukur kemampuan yang seharusnya diukur sesuai dengan tujuan pembelajaran. Dalam konteks evaluasi pembelajaran, validitas butir soal menjadi indikator penting untuk memastikan bahwa instrumen tes memberikan informasi yang akurat mengenai capaian belajar peserta didik. Butir soal yang valid menunjukkan keterkaitan yang kuat antara kinerja peserta didik pada butir tersebut dengan kinerja pada keseluruhan tes.

Dalam penelitian ini, validitas butir soal dianalisis menggunakan uji validitas empiris melalui korelasi antara skor setiap butir soal dengan skor total tes menggunakan rumus Product Moment. Pendekatan validitas empiris dipilih karena penelitian ini bertujuan untuk mengevaluasi kinerja instrumen berdasarkan data hasil uji coba, bukan untuk mengembangkan instrumen baru yang memerlukan validasi ahli secara formal. Dengan demikian, validitas instrumen difokuskan pada kemampuan butir soal dalam merepresentasikan konstruk yang sama dengan tes secara keseluruhan. Penentuan validitas butir soal didasarkan pada perbandingan antara nilai koefisien korelasi (r hitung) dengan nilai r tabel. Dengan jumlah responden sebanyak 30 peserta didik dan taraf signifikansi 5%, diperoleh nilai r tabel sebesar 0,361. Butir soal dinyatakan valid apabila memiliki nilai r hitung ≥ r tabel, sedangkan butir dengan nilai r hitung < r tabel dinyatakan tidak valid, sebagaimana disajikan pada Tabel 4.

Tabel 4. Hasil Analisis Validitas dan Signifiknasi Butir Soal

Nomor Soal	Korelasi/r hitung	Keterangan	Signifikansi
1	0.406	Valid	Signifikan
2	0.392	Valid	Signifikan
3	0.595	Valid	Sangat Signifikan
4	0.654	Valid	Sangat Signifikan
5	0.404	Valid	Signifikan
6	0.327	Tidak Valid	-
7	0.695	Valid	Sangat Signifikan
8	0.716	Valid	Sangat Signifikan
9	0.725	Valid	Sangat Signifikan
10	0.624	Valid	Sangat Signifikan
11	0.595	Valid	Sangat Signifikan
12	0.786	Valid	Sangat Signifikan
13	0.601	Valid	Sangat Signifikan
14	0.604	Valid	Sangat Signifikan
15	0.717	Valid	Sangat Signifikan
16	0.549	Valid	Sangat Signifikan
17	0.552	Valid	Sangat Signifikan
18	0.724	Valid	Sangat Signifikan
19	0.695	Valid	Sangat Signifikan
20	0.701	Valid	Sangat Signifikan
21	0.589	Valid	Sangat Signifikan
22	0.589	Valid	Sangat Signifikan
23	0.862	Valid	Sangat Signifikan
24	0.486	Valid	Sangat Signifikan
25	0.447	Valid	Signifikan
26	0.609	Valid	Sangat Signifikan
27	0.499	Valid	Sangat Signifikan
28	0.763	Valid	Sangat Signifikan
29	0.519	Valid	Sangat Signifikan
30	0.625	Valid	Sangat Signifikan

Hasil analisis validitas butir soal menunjukkan bahwa sebagian besar butir memiliki angka korelasi pada kategori cukup hingga sangat tinggi. Temuan ini mengindikasikan bahwa

mayoritas butir soal telah memiliki keterkaitan yang memadai dengan skor total tes, sebanyak 29 butir dinyatakan valid, dan hanya satu butir soal yang tidak valid yaitu butir nomor 6 dikarenakan nilai korelasi di bawah nilai r tabel, sehingga dapat disimpulkan bahwa mayoritas butir soal mampu mengukur sejauh kemampuan yang seharusnya diukur sesuai dengan tujuan evaluasi pembelajaran.

Selain ditinjau dari besarnya korelasi, validitas butir soal juga dianalisis berdasarkan tingkat signifikansi hubungan antara skor butir dan skor total tes. Hasil analisis menunjukkan bahwa sebagian besar butir soal memiliki korelasi yang signifikan hingga sangat signifikan, yang memperkuat temuan bahwa butir-butir tersebut berkontribusi secara bermakna terhadap pengukuran hasil belajar peserta didik.

Validitas Instrumen

Berdasarkan hasil uji validitas butir soal, dapat disimpulkan bahwa instrumen tes secara keseluruhan memiliki tingkat validitas yang baik. Mayoritas butir soal menunjukkan koefisien korelasi yang signifikan dengan skor total tes, yang menandakan adanya keterkaitan antara kinerja peserta didik pada masing-masing butir dengan kinerja pada keseluruhan tes. Kondisi ini menunjukkan bahwa butir- butir soal yang digunakan telah mampu mengukur konstruk yang sama sesuai dengan tujuan pembelajaran yang ditetapkan.

Dengan dominasi butir soal yang valid, instrumen tes dalam penelitian ini dapat dinyatakan valid secara empiris dan layak digunakan sebagaialat evaluasi hasil belajar peserta didik. Temuan ini juga memperkuat pandangan bahwa instrumen evaluasi yang disusun dan diuji secara sistematis akan menghasilkan data yang lebih akurat dan dapat dijadikan dasar dalam pengambilan keputusan pembelajaran (Hidayati, 2024).

Untuk menginterpretasikan tingkat validitas, maka koefisien kolerasi dikategorikan pada kriteria sebagai berikut, lihat pada tabel 5 (Wordpress, 2015).

Tabel 5. Kategori Koefisien Korelasi

Koefisien Korelasi (r_{xy})	Keterangan
0,800 – 1,000	Korelasi sangat tinggi
0,600 – 0, 800	Korelasi tinggi
0,400 – 0,600	Korelasi cukup
0,200 – 0,400	Korelasi rendah
0,000 – 0,200	Korelasi sangat rendah

Hasil uji validitas instrumen berdasarkan korelasi XY pada data olahan Anates versi 4 *for Windows* menunjukkan koefisien korelasi sebesar 0,94. Ini menginterpretasikan validitas instrumen tes mempunyai korelasi sangat tinggi serta representatif mewakilli indikator.

Reliabilitas Instrumen

Uji reliabilitas dilakukan untuk mengetahui tingkat konsistensi instrumen sebagai alat ukur. Reliabilitas dinyatakan melalui koefisien dengan rentang nilai 0–1, di mana nilai yang semakin mendekati 1 menunjukkan tingkat reliabilitas yang semakin tinggi, sedangkan nilai yang lebih rendah menunjukkan reliabilitas yang rendah (Rahman et al., 2023). Pada penelitian ini, reliabilitas instrumen dianalisis dengan bantuan program ANATES

berdasarkan data hasil uji coba instrumen.

Menurut Guilford, untuk menentukan koefisien reliabilitas alpha croanbach, aturan reliabilitas berikut harus diikuti, lihat pada tabel 6 (Rahman et al., 2023).

Tabel 5. Kategori Koefisien Reliabilitas

Koefisien Reliabilitas	Kriteria
0,80 – 1.00	Reliabilitas sangat tinggi
0.60 – 0,80	Reliabilitas tinggi
0,40 – 0,60	Reliabilitas sedang
0,20 – 0,40	Reliabilitas rendah
0,00 – 0,20	Reliabilitas sangat rendah

Hasil uji reliabilitas pada data olahan Anates versi 4 *for Windows* menunjukkan koefisien sebesar 0,97 yang menunjukkan tingkat reliabilitas sangat tinggi. Nilai reliabilitas tersebut mengindikasikan bahwa butir-butir soal dalam instrumen memiliki konsistensi internal yang sangat baik dan bekerja secara relatif seragam dalam mengukur kemampuan peserta didik. Dengan demikian, instrumen tes yang digunakan dalam penelitian ini dapat dikatakan andal dan layak digunakan sebagai alat evaluasi pembelajaran Al-Qur'an Hadis.

Reliabilitas yang tinggi dalam penelitian ini sejalan dengan hasil penelitian sebelumnya yang menunjukkan bahwa instrumen dengan koefisien reliabilitas tinggi mencerminkan konsistensi internal yang baik dan layak digunakan sebagai alat ukur kemampuan peserta didik. Sehingga hasil pengukuran yang diperoleh dapat dipercaya dan digunakan sebagai dasar pengambilan keputusan dalam evaluasi pembelajaran (Rahman & Khalidi, 2025).

Instrumen evaluasi yang memiliki tingkat validitas dan reliabilitas yang baik menjadi prasyarat penting dalam menghasilkan penilaian yang objektif dan akurat. Tidak hanya pemanfaatan sistem ANATES saja yang menuntut kualitas, sistem penilaian berbasis teknologi, seperti Computer Based Test (CBT), juga menuntut kualitas instrumen yang tinggi agar hasil penilaian dapat dipercaya dan digunakan secara efektif. Penilaian yang didukung oleh instrumen yang berkualitas terbukti mampu meningkatkan objektivitas, efisiensi, serta keakuratan hasil evaluasi pembelajaran (Siraturrahmah, 2025).

KESIMPULAN

Berdasarkan hasil analisis kualitas butir soal mata pelajaran Al-Qur'an Hadis berbasis *Higher Order Thinking Skills* (HOTS) menggunakan program ANATES, dapat disimpulkan bahwa secara umum instrumen tes yang digunakan memiliki kualitas yang baik dan layak digunakan sebagai alat evaluasi pembelajaran. Hasil analisis menunjukkan bahwa mayoritas butir soal berada pada kategori tingkat kesukaran sedang hingga sukar, sehingga instrumen mampu mengukur kemampuan peserta didik secara proporsional dan selaras dengan karakteristik penilaian berbasis *Higher Order Thinking Skills* (HOTS).

Dari aspek daya pembeda, sebagian besar butir soal memiliki kemampuan membedakan yang baik hingga sangat baik antara peserta didik berkemampuan tinggi dan rendah. Analisis fungsi distraktor menunjukkan bahwa secara umum pengecoh telah berfungsi dengan baik.

Hasil uji validitas butir soal menunjukkan bahwa hampir seluruh butir soal termasuk

kategori signifikan dan sangat signifikan terhadap skor total tes, sehingga dinyatakan valid. Hasil uji validitas instrumen juga dikataan valid dengan korelasi sebesar 0,94 yang ini dapat dimaknai bahwa mayoritas instrument representatif mewakili indikator. Sementara itu, hasil uji reliabilitas menunjukkan koefisien reliabilitas sebesar 0,97 yang termasuk dalam kategori sangat tinggi, yang mengindikasikan bahwa instrumen memiliki konsistensi internal yang sangat baik dan dapat dipercaya.

Hasil analisis ini menegaskan pentingnya evaluasi instrumen sebagai bagian dari upaya peningkatan kualitas pembelajaran melalui penyusunan dan perbaikan butir soal yang berkelanjutan. Dengan demikian, instrumen tes HOTS pada mata pelajaran Al-Qur'an Hadis yang dianalisis dalam penelitian ini pada dasarnya layak digunakan sebagai alat evaluasi pembelajaran.

DAFTAR PUSTAKA

Hidayati, H. N. (2024). Sikap mahasiswa terhadap penyelenggaraan evaluasi dosen. *Afeksi: Jurnal Penelitian dan Evaluasi Pendidikan*, 5(6). https://afeksi.id/jurnal/index.php/afeksi

Hidayati, H. N. (2025). Pemanfaatan artificial intelligence pada penilaian pembelajaran mendalam. *Afeksi: Jurnal Penelitian dan Evaluasi Pendidikan*, 6(5). https://afeksi.id/jurnal/index.php/afeksi/article/view/531

Kadir, A. (2015). Menyusun Dan Menganalisis Tes Hasil Belajar Abdul Kadir. Al-Ta'dib, 8(2). https://doi.org/10.31332/atdb.v8i2.411

Kamal, M. J., & Nursikin, M. (2025). Model evaluasi pendidikan nilai di sekolah atau madrasah: Sebuah tawaran model evaluasi komprehensif. *Afeksi: Jurnal Penelitian dan Evaluasi Pendidikan*, 6(5). https://doi.org/10.59698/afeksi.v6i5.551

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. https://doi.org/10.24252/asma.v2i2.16569

Muniroh, L. (2024). Pengembangan instrumen evaluasi pilihan ganda berbasis higher order thinking skills (HOTS). *Jurnal Rumpun Manajemen dan Ekonomi*, 1(3), 676–687.

Purwati, L. M., Arianty, R., Syakilah, D. M., Ridlo, S., & Susilaningsih, E. (2021). Analisis soal tes pilihan ganda berbasis higher order thinking skill menggunakan aplikasi ANATES Windows versi 4.0.9. Jurnal Pendidikan Universitas Garut, 15(2), 460–473. https://doi.org/10.52434/jp.v15i2.1287

Rahman, A. (2025). Analisis kualitas instrumen evaluasi pembelajaran ditinjau dari validitas, reliabilitas, daya pembeda, dan tingkat kesukaran. *Afeksi: Jurnal Penelitian dan Evaluasi Pendidikan*, 6(5).

Rahman, I. A., Viola, M. A., Masita, & Vilanti, F. A. (2023). Uji validitas dan reliabilitas kualitas sarana dan prasarana akademik terhadap prestasi belajar mahasiswa FKIP Universitas Jambi. Jurnal Pendidikan Tambusai, 7(3), 28965–28966. https://doi.org/10.31004/jptam.v7i3.11627

Saputra, H. D., Purwanto, W., Setiawan, D., Fernandez, D., & Putra, R. (2022). Hasil belajar mahasiswa: Analisis butir soal tes. *Edukasi: Jurnal Pendidikan*, 20(1), 15–27.

https://doi.org/10.31571/edukasi.v20i1.3432

Sarnita, F., Adnyana, P. B., & Rapi, N. K. (2025). Uji validitas dan reliabilitas instrumen literasi numerasi materi tata surya untuk siswa tunanetra. *Jurnal Pendidikan Ilmu Pengetahuan Alam (JP-IPA)*, 6(2). https://doi.org/10.56842/jp-ipa

Siraturrahmah, R. M. (2025). Pemanfaatan platform computer based test (CBT) dalam sistem penilaian pembelajaran. *Afeksi: Jurnal Penelitian dan Evaluasi Pendidikan*, 6(5). https://doi.org/10.59698/afeksi.v6i6.702

Sugiyono. (2019). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Bandung: Alfabeta.

Wikipedia. (2024). *Higher-order thinking*. https://en.wikipedia.org/wiki/Higher-order_thinking (diakses 30 Desember 2025).

WordPress (2015). Tak Lelah Belajar. Validitas. https://taklelahbelajar.wordpress.com/2015/02/11/validitas/ (diakses 30 Desember 2025).