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 HOTS pada Mata Pelajaran SKI

Anisa Tri Wulandari [1], Heny Narendrany Hidayati [2]
Universitas Syarif Hidayatullah Jakarta, Indonesia
Email: anisatriwulandari2906@gmail.com , heny.fitk@yahoo.com

Abstrak
Penelitian ini memiliki tujuan untuk menganalisis kualitas butir soal yang berfokus pada Higher Order Thinking Skills (HOTS) pada mata pelajaran Sejarah Kebudayaan Islam (SKI) di MTs Muhammadiyah 1 Ciputat. Penelitian menggunakan metode deskriptif kuantitatif dengan quota sampel, yaitu 31 responden peserta didik kelas VIII. Teknik pengumpulan data menggunakan instrumen tes pilihan ganda 30 butir soal yang diuji coba melalui Google Form. Analisis data menggunakan statistik deskriptif dengan aplikasi Anatest untuk menilai tingkat kesukaran, daya pembeda, fungsi distraktor, validitas butir, dan reliabilitas tes. Hasil penelitian menunjukkan bahwa mayoritas soal memiliki tingkat kesukaran sedang, daya pembeda baik hingga sangat baik, dan fungsi distraktor yang cukup efektif. Sebagian besar butir soal valid, sedangkan reliabilitas tes tergolong tinggi. Walaupun demikian, terdapat beberapa butir soal yang sangat mudah, sangat sukar, atau memiliki daya pembeda negatif serta distraktor yang kurang berfungsi, sehingga perlu direvisi atau diganti. Temuan ini menunjukkan bahwa instrumen tes HOTS SKI secara umum layak digunakan untuk evaluasi pembelajaran, namun perlu perbaikan pada beberapa butir soal untuk meningkatkan kualitas tes.
Kata Kunci: *Analisis Butir Soal, HOTS, Sejarah Kebudayaan Islam, Evaluasi Pembelajaran, Tes Pilihan Ganda*

PENDAHULUAN

Analisis butir soal adalah suatu proses sistematik untuk mengevaluasi kualitas butir-butir soal tes, khususnya tes objektif. Tujuan analisis butir soal adalah memperoleh soal yang bermutu, yaitu soal yang dapat mempresentasikan keadaan yang sebenarnya di lapangan dan memenuhi lima syarat, yaitu taraf sukar, daya beda, fungsi distraktor, validitas, dan reliabilitas serta sudah diuji coba (Lasmy dkk., 2021). Penilaian terhadap hasil belajar merupakan salah satu komponen penting yang harus dilaksanakan dalam proses pembelajaran. Melalui evaluasi, pengajar dapat mengetahui tingkat pencapaian kompetensi peserta didik secara kualitas proses pembelajaran yang telah dilakukan, selain itu guru juga dapat mengevaluasi proses pembelajaran yang telah dilaksanakan di dalam kegiatan belajar mengajar. Evaluasi berfungsi sebagai umpan balik agar siswa tahu sejauh mana pemahamannya terhadap materi pelajaran yang telah diberikan di kelas. Evaluasi tidak hanya dilakukan di akhir pembelajaran tetapi juga dilakukan sepanjang proses pembelajaran. Alat evaluasi yang biasa digunakan di sekolah salah satu adalah tes objektif berbentuk pilihan ganda.

Tes evaluasi dalam pembelajaran harus disusun berdasarkan kisi-kisi pembelajaran yang telah dibuat berdasarkan capaian pembelajaran. Materi yang telah diajarkan, memenuhi kriteria tes yang baik, yaitu valid, reliabel, memiliki tingkat kesukaran yang proporsional, daya pembeda yang baik, serta pengecoh yang berfungsi dengan baik (Arikunto, 2018). Secara teori, kualitas tes sangat ditentukan oleh kualitas setiap butir soal. Butir soal yang baik harus mampu mengukur kompetensi yang seharusnya diukur, membedakan peserta didik yang

berkemampuan tinggi dan rendah, serta memiliki tingkat kesukaran yang seimbang (Arikunto, 2018). Untuk itu, sangat penting butir soal yang dibuat dianalisis terlebih dahulu agar hasil tes memberikan gambaran yang tepat terhadap kemampuan peserta didik.

Adanya tuntutan pembelajaran tingkat tinggi mengharuskan siswa untuk memiliki keterampilan berpikir tingkat tinggi. Untuk itu, keterampilan berpikir tingkat tinggi pembelajaran tidak hanya sebatas mengetahui, memahami, dan mengaplikasikan saja, tetapi juga mampu menganalisis, mengevaluasi dan menciptakan dari hasil proses pembelajaran tersebut (Insani dkk., 2023). Berdasarkan hasil uji coba kelapangan berupa analisis butir soal pada tes pilihan ganda tingkat HOTS pada mata pelajaran SKI materi Dinasti Abbasiyah yang diujikan kepada 31 peserta didik kelas VIII di MTs Muhammadiyah 1 Ciputat dengan 30 butir soal, ditemukan bahwa meski secara umum nilai rata-rata berada pada kategori sedang, masih terdapat beberapa butir soal yang memiliki daya pembeda rendah atau negatif, tingkat kesukaran terlalu mudah atau terlalu sukar, serta validitas butir yang tidak signifikan. Temuan tersebut menunjukkan bahwa instrumen tes yang digunakan belum sepenuhnya memenuhi kriteria instrumen evaluasi yang berkualitas.

Pengukuran tes hasil belajar yang akurat memerlukan instrumen evaluasi yang sahih dan andal. Dalam teori evaluasi pendidikan, analisis butir soal merupakan prosedur sistematis yang dilakukan untuk mengetahui karakteristik setiap butir soal berdasarkan data empiris hasil pengerjaan peserta didik (Sudijono, 2015). Analisis ini mencakup pengukuran tingkat kesukaran, daya pembeda, fungsi distraktor, validitas, dan reliabilitas butir soal. Tingkat kesukaran menunjukkan proporsi peserta didik yang dapat menjawab suatu soal dengan benar, daya pembeda menunjukkan kemampuan soal dalam membedakan peserta didik berkemampuan tinggi dan rendah, sedangkan validitas butir menunjukkan sejauh mana suatu butir soal, mengukur apa yang seharusnya diukur (Arikunto, 2019). Pengecoh dikatakan baik bila mampu menarik peserta didik yang kurang menguasai materi untuk memilihnya, sehingga memperkuat fungsi diskriminatif soal.

Urgensi penelitian ini adalah kebutuhan guru dan satuan pendidikan untuk instrumen evaluasi yang mampu memberikan informasi yang akurat mengenai hasil belajar peserta didik. Instrumen tes yang tidak dianalisis berpotensi menghasilkan keputusan yang kurang tepat, seperti dalam penentuan ketuntasan belajar, pemberian remedial, maupun pengayaan. Oleh karena itu, analisis butir soal perlu dilakukan sebagai bagian dari peningkatan mutu evaluasi pembelajaran dan profesionalisme guru (Arifin, 2017).

Berdasarkan latar belakang dan dasar pemikiran yang telah dikemukakan, permasalahan dalam penelitian ini berfokus pada kualitas butir soal tes pilihan ganda yang digunakan sebagai alat evaluasi pembelajaran, permasalahan tersebut mencakup bagaimana tingkat kesukaran setiap butir soal, bagaimana daya pembeda soal dalam membedakan peserta didik yang berkemampuan tinggi dan rendah, serta bagaimana kualitas pengecoh pada setiap alternatif jawaban dalam tes pilihan ganda tersebut. Selain itu, penelitian ini juga menelaah bagaimana validitas butir soal yang ditinjau dari korelasi antara skor butir dengan skor total tes dan menelaah bagaimana reliabilitas instrumen tes tersebut. Menentukan kelayakan setiap butir soal, apakah layak digunakan kembali, perlu direvisi, atau harus dibuang sebagai instrumen evaluasi pembelajaran. Penelitian ini bertujuan untuk menganalisis kualitas butir

soal tes pilihan ganda sebagai instrumen evaluasi pembelajaran ditinjau dari tingkat kesukaran, daya pembeda, kualitas pengecoh, validitas butir, dan reliabilitas tes, serta menentukan kelayakan setiap butir soal untuk digunakan kembali, direvisi, atau dibuang.

METODE

Penelitian ini menggunakan metode penelitian deskriptif kuantitatif. Pendekatan kuantitatif merupakan suatu proses penelitian yang melibatkan perumusan hipotesis atau prediksi hasil, pengumpulan data empiris, analisis data, dan penarikan kesimpulan, dengan menggunakan aspek pengukuran, perhitungan, rumus, dan data numerik, serta menerapkan metode statistik untuk mengolah data (Rukminingsih dkk., 2020). Untuk menelaah kualitas butir soal tes pilihan ganda yang digunakan sebagai alat evaluasi pembelajaran, deskriptif kuantitatif ini dipilih karena memungkinkan peneliti memperoleh gambaran numerik mengenai tingkat kesukaran, daya pembeda, kualitas pengecoh, validitas butir, dan reliabilitas instrumen secara sistematis. Populasi penelitian terdiri dari seluruh peserta didik kelas VIII di MTs Muhammadiyah 1 Ciputat yang dilakukan pada Selasa 25 November 2025. Untuk memperoleh data yang komprehensif, digunakan teknik pengambilan data menggunakan quota sampel dengan jumlah 31 responden tes.

Prosedur penelitian dilakukan melalui beberapa tahap. Pertama, peneliti menyusun butir soal pilihan ganda berdasarkan kisi-kisi yang telah ditentukan. Selanjutnya, teknik pengumpulan data menggunakan instrumen tes berupa soal pilihan ganda yang dibagikan melalui Google Form diberikan kepada seluruh peserta didik, kemudian hasil tes yang terekam secara otomatis diunduh lalu dianalisis menggunakan aplikasi Anatest. Analisis meliputi perhitungan skor, tingkat kesukaran, daya pembeda, kualitas pengecoh, validitas butir, dan reliabilitas tes.

Instrumen penelitian berupa tes pilihan ganda yang dirancang untuk mengukur pencapaian kompetensi peserta didik sesuai indikator pembelajaran. Konstruk yang diukur meliputi kemampuan kognitif peserta didik dalam menyelesaikan soal sesuai materi pembelajaran. Validitas instrumen ditentukan melalui korelasi antara skor tiap butir dengan skor total tes, butir yang memiliki korelasi signifikan dianggap valid. Sementara itu, reliabilitas tes dihitung menggunakan koefisien reliabilitas *Kuder Richardson* untuk mengetahui konsistensi internal instrumen.

Data penelitian dikumpulkan melalui Google Form serta data dianalisis menggunakan statistik deskriptif dengan program statistik seperti Anatest. Analisis meliputi penghitungan tingkat kesukaran setiap butir soal, daya pembeda antara peserta didik berkemampuan tinggi dan rendah, kualitas pengecoh pada alternatif jawaban, validitas butir, serta reliabilitas tes secara keseluruhan. Hasil analisis ini digunakan untuk menentukan kelayakan setiap butir soal, apakah dapat digunakan kembali, perlu direvisi, atau harus dibuang sebagai instrumen evaluasi pembelajaran.

HASIL DAN PEMBAHASAN

Penilaian merupakan komponen penting dalam proses pembelajaran, sebab melalui penilaian dapat diketahui sejauh mana tujuan pembelajaran telah tercapai (Hidayati, 2025). Tes yang dianalisis terdiri atas 30 butir soal pilihan ganda dengan jumlah peserta didik

sebanyak 31 orang. Tes dilakukan pada mata pelajaran SKI dengan tingkat soal HOTS yaitu level kognitif C4, C5, dan C6. Pemberian skor menggunakan sistem dikotomi, yaitu skor 1 untuk jawaban benar dan skor 0 untuk jawaban salah, sehingga skor maksimum yang dapat diperoleh peserta didik adalah 30, dan skor minimum adalah 0. Hasil pengolahan data menunjukkan bahwa rata-rata skor tes adalah 17,00, dengan standar deviasi sebesar 5,82. Nilai rata-rata tersebut berada di sekitar 56,7% dari skor maksimum, yang menunjukkan bahwa secara umum tingkat penguasaan materi peserta didik berada pada kategori sedang. Standar deviasi yang relatif besar menunjukkan adanya variasi kemampuan yang cukup lebar antar peserta didik. Kondisi ini mengindikasikan bahwa tes mampu menangkap perbedaan kemampuan individu, sehingga layak untuk dianalisis lebih lanjut dari sisi kualitas butir soal (Sudijono, 2015).

Taraf Sukar

Taraf sukar setiap soal berbeda satu sama lain. Soal-soal dengan tingkat *Higher Order Thinking Skills* (HOTS) belum tentu butir-butir soal tersebut juga mempunyai taraf sukar yang tinggi. Kemampuan berpikir tingkat tinggi dapat dilatih dalam kegiatan belajar mengajar di kelas, yaitu dengan menyediakan kesempatan kepada peserta didik untuk menemukan konsep pengetahuan berbasis aktivitas (Hidayati, n.d.). Tingkat kesukaran soal dapat diketahui melalui proporsi jawaban testee yang menjawab benar dari butir-butir soal tes tersebut. Tingkat kesukaran butir soal adalah bagian penting dalam pengujian tes atau kemampuan. Tingkat kesukaran soal dapat berkontribusi pada pengembangan soal yang lebih efektif dan berkualitas tinggi (Rahman & Khalid, 2015).

Tabel 1. Kriteria Tingkat Kesukaran Butir Soal

Kriteria Tingkat Kesukaran	Klasifikasi
0, 00 – 0, 25	Sukar
0, 25 – 0, 75	Sedang
0, 75 – 1, 00	Mudah

Tingkat kesukaran soal dipengaruhi oleh kemampuan peserta didik. Idealnya, tingkat kesukaran soal hasil belajar sekitar 0, 50, tingkat kesukaran keseluruhan didapat dari penjumlahan semua butir soal (Hidayati, 2007). Semakin besar nilai P maka semakin rendah tingkat kesukaran soal atau soal itu makin mudah. Jika tingkat kesukaran mendekati 1, 00 berati semua peserta tes dapat menjawab butir soal itu secara benar. Sebaliknya jika tingkat kesukaran mendekati 0, 00 berarti sedikit dari peserta tes yang menjawab soal dengan benar.

Tabel 2. Hasil Uji Tingkat Kesukaran Soal

No	Nomor Soal	Rentang Tingkat Kesukaran	Kategori
1	1, 25	0, 7097 dan 0, 9355	Sangat mudah
2	8, 17, 23, 24, 26, 30	0, 7419 – 0, 8387	Mudah
3	2, 3, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 22, 28, 29	0, 3226 – 0, 6774	Sedang
4	21, 27	0, 1613 - 0, 2903	Sukar
5	4	0, 1290	Sangat sukar

Berdasarkan tabel 2 di atas, terdapat 2 soal sangat mudah, 6 soal mudah, 19 soal sedang, 2 soal sukar, dan 1 soal sangat sukar. Soal-soal dengan kategori sedang masih dapat digunakan dalam tes-tes hasil belajar yang akan datang, soal-soal kategori mudah dapat direvisi sehingga benar-benar mampu mengukur kompetensi peserta didik, sedangkan untuk soal-soal dengan tingkat sangat mudah dan sangat sukar harus diganti atau dibuang. Soal kategori sukar hingga sangat sukar perlu dikaji ulang untuk dapat diketahui penyebab banyak peserta didik tidak mampu menjawab soal tersebut dengan benar. Soal kategori sukar hingga sangat sukar dapat digunakan pada tes dengan tingkat kelulusan tinggi. Sementara itu, untuk soal kategori sangat mudah hingga mudah juga dapat dikaji ulang dan diteliti penyebab soal tersebut mudah dijawab oleh peserta didik dan soal-soal tersebut dapat digunakan pada tes yang berfungsi hanya sebagai formalitas saja.

Daya Pembeda

Daya pembeda merupakan kemampuan suatu butir soal untuk membedakan peserta didik yang memiliki kemampuan tinggi (kelompok unggul) dan kemampuan rendah (kelompok asor). Analisis daya pembeda dilakukan dengan membandingkan jumlah jawaban benar pada kelompok atas dan kelompok bawah.

Daya pembeda yang baik adalah yang mampu membedakan antara kelompok atas dan kelompok bawah. Indeks daya pembeda soal berkisar antara -1 hingga +1. Indeks daya pembeda bertanda positif menunjukkan bahwa butir soal tersebut sudah memiliki daya pembeda, artinya testee berkemampuan tinggi lebih banyak dapat menjawab betul terhadap butir soal tersebut, sedangkan testee berkemampuan rendah lebih banyak menjawab salah. Jika indeks daya pembeda 0, 00, maka butir soal tidak memiliki daya pembeda sama sekali. Jika indeks daya pembeda bertanda negatif, maka butir soal tersebut lebih banyak dijawab benar oleh testee kelompok berkemampuan rendah dari pada kelompok atas (Hidayati, n.d.).

Tabel 3. Indeks Daya Pembeda

No	Indeks Daya Pembeda	Klasifikasi	Interpretasi
1	Kurang dari 0, 20	Poor	Daya pembeda lemah
2	0, 20 – 0, 39	Satisfactory	Daya pembeda yang cukup (sedang)
3	0, 40 – 0, 69	Good	Daya pembeda yang baik
4	0, 70 – 1,00	Excellent	Daya pembeda yang baik sekali
5	Bertanda negatif		Daya pembeda negatif (jelek sekali)

Daya pembeda suatu item dapat diidentifikasi melalui indeks diskriminasi, yang menunjukkan kemampuan item membedakan antara kelompok atas dan kelompok bawah. Perhitungan didasarkan pada pembagian peserta tes menjadi kelompok atas dan kelompok bawah untuk menilai daya pembeda diskriminasi item tersebut (Lasmy dkk., 2021). Setelah hasil uji coba didapati dan skor telah diurutkan dari yang paling tinggi hingga paling rendah, maka didapat dibagi menjadi 2 kelompok yaitu kelompok atas dan kelompok bawah. Terdapat 8 testee berada pada kategori kelompok atas dengan skor tertinggi 27 menjawab soal benar dan 3 soal dijawab salah, sedangkan untuk kelompok bawah terdapat 8 testee juga, di mana skor terendah 5 soal dijawab benar dan 25 soal dijawab salah.

Tabel 4. Daya Pembeda Butir Soal

Kategori Daya Pembeda	Rentang DP	Nomor Butir
Sangat Baik	≥ 0,70	2, 8, 10, 11, 18, 22, 23
Baik	0,40 – 0,69	1, 3, 6, 7, 9, 12, 14, 16, 19, 20, 24, 26, 28, 29
Cukup	0,20 – 0,39	13, 15, 17, 30
Buruk	0,00 – 0,19	4, 5, 21
Negatif	< 0,00	27

Berdasarkan tabel tersebut, dapat diketahui bahwa mayoritas butir soal memiliki daya pembeda baik hingga sangat baik. Hal ini menunjukkan bahwa sebagian besar butir mampu membedakan peserta didik berkemampuan tinggi dan rendah secara efektif. Sebaliknya, butir nomor 21 memiliki daya pembeda nol, sedangkan butir nomor 27 memiliki daya pembeda negatif (-0, 12). Butir dengan daya pembeda negatif menunjukkan adanya ketidakwajaran, di mana kelompok bawah lebih banyak menjawab benar dibandingkan kelompok atas. Butir seperti ini sebaiknya dibuang atau direvisi secara menyeluruh.

Fungsi Distraktor (Pengecoh)

Pilihan jawaban pada tes objektif biasanya berjumlah 3-5 buah, di mana ada 1 dari sekian opsi jawaban yang benar dan sisanya merupakan jawaban yang salah (Hidayati, 2007). Distraktor atau pengecoh adalah alternatif jawaban selain kunci yang berfungsi untuk mengecoh peserta didik yang belum menguasai materi. Distraktor yang baik adalah distraktor yang dipilih oleh peserta didik kelompok bawah dan relatif tidak dipilih oleh kelompok atas. Distraktor berfungsi jika sekurang-kurangnya dipilih oleh testee dan distraktor dipilih lebih banyak oleh kelompok bawah.

Hasil analisis kualitas distraktor menunjukkan bahwa secara umum pengecoh berfungsi dengan baik, ditandai dengan banyaknya pengecoh yang berada pada kategori baik (+) dan sangat baik (++). Namun, pada beberapa butir soal ditemukan distraktor dengan kategori buruk (–), sangat buruk (– –), dan sangat tidak berfungsi (– – –), terutama pada butir yang tergolong sangat mudah dan sangat sukar. Kondisi ini menunjukkan bahwa alternatif jawaban pada butir tersebut kurang homogen atau terlalu mencolok sehingga mudah dieliminasi oleh peserta didik.

Distraktor yang tidak berfungsi sebaiknya diperbaiki dengan cara menyusun alternatif jawaban yang lebih logis, sejajar, dan sesuai dengan miskonsepsi yang sering dialami peserta didik. Berdasarkan hasil uji kualitas pengecoh pada instrumen tes Try Out mata pelajaran Sejarah Kebudayaan Islam (SKI), diketahui bahwa fungsi distraktor dapat diklasifikasikan ke dalam beberapa kriteria, yaitu sangat baik, baik, kurang baik, buruk, dan sangat buruk. Pengelompokan ini didasarkan pada banyaknya siswa yang memilih setiap distraktor serta pola pemilihannya.

Tabel 5. Hasil Uji Fungsi Dsitraktor

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

Berdasarkan tabel tersebut, dapat diketahui bahwa sebagian besar butir soal memiliki distraktor dengan kategori baik hingga sangat baik. Hal ini menunjukkan bahwa pilihan jawaban yang disediakan sudah cukup homogen dan mampu mengecoh siswa yang belum menguasai materi. Distraktor pada butir soal tersebut berfungsi dengan baik karena dipilih oleh siswa dan tidak menimbulkan pola jawaban yang terlalu mencolok.

Namun, masih terdapat beberapa butir soal yang memiliki distraktor kurang baik hingga sangat buruk. Distraktor pada butir soal nomor 4, 13, 25, dan 27 menunjukkan kualitas yang rendah karena tidak dipilih oleh siswa atau terlalu jelas kesalahannya. Kondisi ini menyebabkan soal menjadi kurang efektif dan berpotensi menurunkan daya pembeda tes. Oleh karena itu, butir soal dengan distraktor kurang baik, buruk, dan sangat buruk perlu direvisi dengan memperbaiki redaksi pilihan jawaban agar lebih logis dan seimbang. Dengan demikian, dapat disimpulkan bahwa secara umum fungsi distraktor pada instrumen tes Try Out SKI sudah berjalan dengan baik, meskipun masih diperlukan perbaikan pada beberapa butir soal agar kualitas tes semakin optimal.

Validitas Butir Soal

Validitas dari suatu tes adalah ketepatan mengukur yang dimiliki oleh sebuah item atau butir yang merupakan bagian tak terpisahkan dari tes sebagai suatu totalitas dalam mengukur apa yang seharusnya diukur lewat butir soal tersebut (Hidayati, 2007). Validitas butir soal menunjukkan sejauh mana suatu butir soal mampu mengukur kemampuan yang seharusnya diukur sesuai dengan tujuan pembelajaran. Dalam penelitian ini, validitas butir soal dianalisis melalui korelasi Product Moment antara skor butir soal dengan skor total tes. Validitas merupakan ketepatan dan kecermatan suatu instrumen dalam mengukur apa yang seharusnya diukur, mencakup validitas isi (mengukur konten materi), validitas konstruk (mengukur konsep atau teoritis), dan validitas empiris atau kriteria (berdasarkan kriteria internal atau eksternal) (Ramadhan dkk., 2024). Validitas isi ditentukan melalui analisis logis terhadap kisi-kisi tes, validitas konstruk melalui penelaahan teoritis dan justifikasi ahli, sedangkan validitas empiris ditentukan berdasarkan uji coba dengan kriteria internal (validitas butir) atau eksternal (Ramadhan dkk., 2024).

Validitas butir soal dalam penelitian ii didasarkan pada tabel r Product Moment. Dengan jumlah responden sebanyak 30 peserta didik dan taraf signifikasi 5%, diperoleh nilai r tabel = 0, 361. Butir soal dinyatakan valid jika nilai r hitung ≥ r tabel (0, 361), sedangkan butir soal dengan r hitung < 0, 361 dinyatakan tidak valid.

Tabel 6. Klasifikasi Koefisien Validitas Tes

Kriteria Skala	Klasifikasi
0 – 0, 20	Sangat rendah
0, 21 – 0, 40	Rendah
0, 41 – 0, 60	Cukup
0, 61 – 0, 80	Tinggi
0, 81 – 1, 00	Sangat Tinggi
	(Novia dkk., 2020)

Hasil analisis menunjukkan bahwa sebagian besar butir soal memiliki koefisien korelasi yang signifikan dan sangat signifikan. Hal ini berarti butir soal tersebut telah valid dan mampu mencerminkan kemampuan siswa secara keseluruhan. Akan tetapi, terdapat beberapa butir soal yang memiliki koefisien korelasi rendah dan tidak signifikan, bahkan terdapat satu butir soal yang menunjukkan korelasi negatif. Butir soal dengan korelasi negatif menandakan bahwa soal tersebut tidak sejalan dengan tujuan pengukuran dan dapat mengganggu keakuratan hasil tes. Jika berdasarkan tabel 6. kriteria koefisien validitas tes, dari hasil uji coba maka didapatkan:

Tabel 7. Hasil Analisis Validitas Butir Soal

Nomor Soal	Rentang Validitas Butir	Keterangan
27	-0, 123	Negatif
5, 21, 30	0,137 - 0,199	Sangat rendah
4, 7, 9, 12, 13, 14, 15, 17, 20	0, 260 – 0, 374	Rendah
1, 3, 6, 11, 18, 19, 24, 25, 26, 28, 29	0, 440 – 0, 548	Cukup
2, 8, 10, 22, 23	0, 618 – 0, 683	Tinggi

Tabel 8. Signifikansi Hasil Uji Validitas Butir Soal

Nomor Soal	Signifikansi
1, 2, 3, 8, 10, 11, 16, 18, 19, 22, 23, 24, 25, 26, 28, 29	Sangat signifikan
6, 7, 12, 14	Signifikan
4, 5, 9, 13, 15, 17, 20, 21, 27, 30	--

Berdasarkan tabel 8 di atas, dapat disimpulkan bahwa sebagian besar butir soal telah memenuhi kriteria validitas, sedangkan butir soal yang tidak valid perlu direvisi atau dikeluarkan dari perangkat tes. Validitas butir soal dianalisis melalui korelasi antara skor butir

dengan skor total. Nilai koefisien korelasi tersebut memenuhi kriteria validitas berdasarkan r tabel Product Moment sebesar 0, 361. Hasil uji validitas menunjukkan dari 30 butir soal, 20 butir soal memiliki korelasi signifikan hingga sangat signifikan dengan nilai r hitung ≥ 0, 361 sehingga dinyatakan valid dan 10 butir soal memiliki korelasi tidak signifikan dengan nilai r hitung < 0, 361 sehingga dinyatakan kurang valid. Butir soal yang valid menunjukkan bahwa butir tersebut mampu mengukur kemampuan yang sama dengan keseluruhan tes. Sebaliknya, butir yang tidak valid menunjukkan bahwa butir tersebut kurang memberikan kontribusi terhadap pengukuran hasil belajar peserta didik. Butir nomor 4, 5, 13, 15, 21, dan 27 secara konsisten menunjukkan hasil yang kurang baik pada analisis taraf kesukaran, daya pembeda, dan validitas, sehingga perlu menjadi prioritas dalam proses revisi atau penghapusan. Sedangkan validitas instrumen secara keseluruhan atau korelasi XY untuk hasil uji didapatkan nilai sebesar 0, 73. Untuk itu validitas instrumen tersebut berada pada kategori validitas tinggi.

Reliabilitas

Uji reliabilitas dilakukan untuk mengetahui sejauh mana tingkat kepercayaan atau konsistensi butir soal dalam mengukur sehingga dapat ditentukan apakah tes hasil belajar yang disusun telah memiliki daya keajegan atau kepercayaan yang tinggi, sehingga instrumen tes hasil belajar tersebut dapat dikatakan reliabel. Reliabel adalah tingkat konsistensi hasil ukur atau dapat mengukur keadaan yang sebenarnya dari keadaan peserta didik yang diukur (Hidayati, n.d.). Uji reliabilitas tes objektif, yaitu dengan menggunakan rumus *Kuder Richardson*. Analisis butir soal menggunakan rumus *Kuder Richardson* (KR-20) untuk data nominal, dengan rentang reliabilitas 0, 0 – 1, 0. Metode ini cocok digunakan untuk soal objektif (Muniroh, 2024).

Reliabilitas tes menunjukkan tingkat konsistensi suatu instrumen dalam mengukur hasil belajar siswa. Tes yang reliabel akan memberikan hasil yang relatif sama apabila digunakan berulang kali pada kondisi yang setara. Pada penelitian ini, reliabilitas tes dianalisis menggunakan metode belah dua (split-half) dengan membandingkan skor soal ganjil dan genap (Sugiyono, 2013). Hasil analisis menunjukkan bahwa nilai korelasi antara skor ganjil dan genap sebesar 0,73, sedangkan koefisien reliabilitas tes sebesar 0,84. Nilai tersebut menunjukkan bahwa tes memiliki reliabilitas tinggi. Dengan demikian, instrumen tes Try Out SKI dapat dikatakan konsisten dan dapat dipercaya sebagai alat evaluasi hasil belajar siswa.

Tabel 9. Kriteria Derajat Reliabilitas Butir Soal

Kriteria	Klasifikasi
< 0, 20	Tidak reliabel
0, 20 - 0, 40	Reliabilitas rendah
0, 40 - 0, 60	Reliabilitas sedang
0, 60 - 0, 80	Reliabilitas tinggi
0, 80 - 1, 00	Reliabilitas sangat tinggi
	(Muniroh, 2024)

Tabel 10. Data Reliabilitas Hasil Uji

Komponen	Nilai
Jumlah Siswa	31
Jumlah Butir Soal	30
Rata-rata Skor	17, 00
Simpangan Baku	5, 82
Korelasi Ganjil–Genap	0, 73
Koefisien Reliabilitas	0, 84
Kriteria	Tinggi

Reliabilitas tes menunjukkan tingkat keajekan atau konsistensi suatu instrumen dalam mengukur kemampuan. Hasil analisis menunjukkan bahwa koefisien reliabilitas tes sebesar 0, 84, yang berada pada kategori sangat tinggi. Nilai reliabilitas ini menunjukkan bahwa tes memiliki konsistensi internal yang baik dan dapat dipercaya sebagai alat ukur hasil belajar. Selain itu, korelasi antara skor ganjil dan skor genap sebesar 0,73, yang menunjukkan hubungan yang kuat antara dua belahan tes. Dengan demikian, tes yang dianalisis dapat dikatakan andal dan layak digunakan, baik untuk evaluasi formatif maupun sumatif. Reliabilitas yang tinggi mengindikasikan bahwa butir-butir soal telah dikonstruksi dengan baik, mempertimbangkan aspek bahasa, struktur, dan konteks lokal (Setiyadi dkk., 2025).

KESIMPULAN

Berdasarkan hasil analisis butir soal tes pilihan ganda mata pelajaran Sejarah Kebudayaan Islam (SKI) materi Dinasti Abbasiyah, dapat disimpulkan bahwa secara umum instrumen tes yang digunakan telah memiliki kualitas cukup baik. Sebagian besar butir soal berada pada tingkat kesukaran sedang, memiliki daya pembeda baik hingga sangat baik, serta distraktor yang berfungsi dengan baik. Selain itu, mayoritas butir soal menunjukkan validitas yang signifikan, sehingga mampu mengukur kemampuan peserta didik sesuai dengan tujuan pembelajaran. Namun demikian, masih ditemukan beberapa butir soal dengan tingkat kesukaran sangat mudah atau sangat sukar, daya pembeda rendah atau negatif, distraktor tidak berfungsi, serta validitas rendah, yang menunjukkan bahwa instrumen tes belum sepenuhnya optimal.

Ditinjau dari aspek reliabilitas, tes memiliki koefisien reliabilitas sebesar 0,84 yang berada pada kategori sangat tinggi. Hal ini menunjukkan bahwa instrumen tes memiliki konsistensi internal yang baik dan dapat dipercaya sebagai alat evaluasi hasil belajar peserta didik. Meskipun demikian, keberadaan beberapa butir soal yang tidak valid dan memiliki daya pembeda negatif berpotensi mengurangi ketepatan pengukuran apabila tidak dilakukan perbaikan. Oleh karena itu, butir soal yang berkualitas baik dapat digunakan kembali, sedangkan butir soal yang kurang baik perlu direvisi atau dibuang agar instrumen evaluasi yang digunakan benar-benar mencerminkan kemampuan peserta didik secara akurat.

Berdasarkan temuan tersebut, disarankan agar guru melakukan analisis butir soal secara rutin sebagai bagian dari proses evaluasi pembelajaran. Hasil analisis dapat dimanfaatkan untuk memperbaiki kualitas soal, khususnya dalam penyusunan soal HOTS yang seimbang dari segi tingkat kesukaran, daya pembeda, dan fungsi distraktor. Bagi peneliti selanjutnya,

penelitian ini dapat dijadikan rujukan untuk mengembangkan instrumen evaluasi yang lebih komprehensif dengan melibatkan jumlah responden yang lebih besar serta mengombinasikan analisis kuantitatif dengan telaah kualitatif oleh ahli, sehingga mutu evaluasi pembelajaran dapat terus ditingkatkan.

DAFTAR PUSTAKA

Arifin, Z. (2017). *Evaluasi Pembelajaran*. Bandung: Remaja Rosdakarya.

Arikunto, S. (2018). *Dasar-dasar Evaluasi Pendidikan* (Edisi Revisi). Jakarta: Bumi Aksara.

Arikunto, S. (2019). *Prosedur Penelitian: Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta.

Hidayati, H. N. (2007). Analisis Perangkat Soal Mata Kuliah Kesehatan Mental. *Tahdzib: Jurnal Pendidikan Agama Islam*, 1(1).

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

Hidayati, H. N. (n.d.). Evaluasi Pembelajaran. *FITK UIN Syarif Hidayatullah Jakarta.*

Insani, Y. W., Tahir, M., & Hasnawati. (2023). Analisis Kesulitan Guru Menyusun Soal Berbasis Higher Order Thingking Skills (HOTS) pada Muatan Materi IPS di SDN 03 Jembatan Gantung. *Jurnal Ilmiah Profesi Pendidikan*, *8*(3), 1719–1724. https://doi.org/10.29303/jipp.v8i3.1561

Lasmy., Yusrizal., & Razali. (2021). Analisis Butir Soal Ujian Sekolah Berstandar Nasional di SMA Kabupaten Aceh Barat. *Jurnal Mudarrisuna: Media Kajian Pendidikan Agama Islam*, *11*(3), 444. https://doi.org/10.22373/jm.v11i3.5500

Muniroh, L. (2024). Pengembangan Instrumen Evaluasi Pilihan Ganda Berbasis HOTS pada Mata Pelajaran. *Jurnal Rumpun Manajemen dan Ekonomi*, *1*(3), 676–687. https://doi.org/10.61722/jrme.v1i3.2235

Novia, T., Wardani, A., Canda., Nurdi., Nurmasyitah. (2020). Analisis Validitas dan Reliabilitas Butir Soal UTS Fisika Kelas X SMA Swasta Muhammadiyah 4 Langsa. *Gravitasi: Jurnal Pendidikan Fisika dan Sains*, (3)1. https://ejurnalunsam.id/index.php/JPFS

Rahman, A., & Khalidi, A. (2025). Analisis Butir Soal Ulangan Akhir Semester Ganjil Pembelajaran IPA Kelas 6 MI Ummul Quraamuntai. *Pediaqu: Jurnal Pendidikan Sosial dan Humaniora*, 4(2).

Ramadhan, F. M., Siroj, R., Afgani, M. W. (2024). Validitas and Reliabilitas. *Journal on Education*, *06*(02), 10967–10975.

Rukminingsih., Adnan, G., & Latief, M. A. (2020). *Metode Penelitian Pendidikan* Penelitian Kuantitatif, Penelitian Kualitatif, Penelitian Tindakan Kelas. Yogyakarta: Erhaka Utama. www.erhakautama.com

Setiyadi, D., Suharini, E., Widitatmoko, A. (2025). Analisis Butir Soal Uraian Bernuansa Etnososial pada Materi IPS Kelas V Sekolah Dasar. *Fondatia.* https://doi.org/10.36088/fondatia.v9i2.5711

Sudijono, A. (2015). *Pengantar Evaluasi Pendidikan*. Jakarta: Rajagrafindo Persada.

Sugiyono. (2013). *Metode Penelitian Kuantitatif, Kualitatif dan R & D*. Bandung: Alfabeta.