

ANALISIS BUTIR SOAL HOTS UNTUK MENINGKATKAN KUALITAS INSTRUMEN PENILAIAN PEMBELAJARAN

Dona Melianda Fitri¹, Luqman Hakim², Vivi Pratiwi³, Kisyia Putri Oktavia⁴, Manisa Eka Sujiarsih⁵, Adila Embarisa Br Brahmana⁶, Ellen Solagracia Sihite⁷, Zufaira Zahra Isna Putri⁸, Bagus Satrio Wibowo⁹
^{1,2,3,4,5,6,7,8,9}Universitas Negeri Surabaya
e-mail: 24080304054@mhs.unesa.ac.id

ABSTRAK

Penelitian ini bertujuan untuk menganalisis kualitas butir soal pilihan ganda berbasis Higher Order Thinking Skills (HOTS) pada mata pelajaran Ekonomi Bisnis dan Administrasi Umum kelas XI menggunakan software ANATES. Analisis dilakukan berdasarkan empat komponen utama, yaitu validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. Penelitian menggunakan pendekatan deskriptif kuantitatif dengan subjek 20 siswa. Data dikumpulkan melalui tes berbasis indikator kompetensi kurikulum. Hasil penelitian menunjukkan bahwa pada aspek validitas, sebagian besar butir berada pada kategori rendah (33,33%) dan cukup (30,56%), sementara kategori sangat tinggi dan sangat rendah masing-masing hanya mencapai 8,33%. Reliabilitas instrumen berada pada rentang cukup hingga sangat tinggi, dengan nilai tertinggi pada indikator 3 sebesar 0,94 dan terendah pada indikator 7 sebesar 0,45. Tingkat kesukaran memperlihatkan distribusi yang dominan pada kategori sedang (66,67%), diikuti kategori sukar (16,67%), mudah (9,72%), dan sangat mudah (6,94%). Pada aspek daya pembeda, kategori baik mendominasi sebanyak 54,17% dan sangat baik 30,56%, meskipun terdapat 4,17% butir soal dengan daya pembeda jelek yang perlu direvisi. Secara keseluruhan, instrumen soal telah memenuhi kriteria kelayakan, namun masih diperlukan perbaikan pada butir dengan validitas dan daya pembeda rendah untuk meningkatkan kualitas asesmen berbasis HOTS.

Kata Kunci: *Administrasi Umum, Analisis butir soal, Anates, Ekonomi Bisnis, HOTS*

ABSTRACT

This study aims to analyze the quality of multiple-choice test items based on Higher Order Thinking Skills (HOTS) in the subjects of Business Economics and Public Administration for 11th-grade students using ANATES software. The analysis was conducted based on four main components: validity, reliability, difficulty level, and discriminating power. The research employed a quantitative descriptive approach with 20 students as subjects. Data were collected through tests developed based on curriculum competency indicators. The results indicate that in terms of validity, most items were classified as low (33.33%) and moderate (30.56%), while very high and very low categories each accounted for only 8.33%. Instrument reliability ranged from moderate to very high, with the highest value of 0.94 on indicator 3 and the lowest value of 0.45 on indicator 7. The difficulty level distribution was dominated by the moderate category (66.67%), followed by difficult (16.67%), easy (9.72%), and very easy (6.94%). Regarding discriminating power, the majority of items fell into the good category (54.17%) and very good (30.56%), although 4.17% of the items had poor discrimination and required revision. Overall, the test instrument met the eligibility criteria; however, improvements are still necessary for items with low validity and low discriminating power to enhance the quality of HOTS-based assessment.

Keywords: *Public Administration, Item Analysis, ANATES, Business Economics, HOTS*

PENDAHULUAN

Evaluasi hasil belajar menempati posisi yang sangat sentral dan krusial dalam keseluruhan ekosistem pendidikan, berfungsi sebagai barometer utama untuk menentukan sejauh mana ketercapaian kompetensi peserta didik telah terwujud. Lebih dari sekadar pemberian angka atau predikat kelulusan, evaluasi merupakan mekanisme umpan balik yang strategis untuk mengukur efektivitas proses pembelajaran yang telah dilaksanakan oleh guru di sekolah. Instrumen penilaian yang berkualitas tinggi tidak hanya berperan sebagai alat ukur kemampuan akademik semata, melainkan juga berfungsi sebagai landasan objektif bagi pengambilan keputusan pedagogis, memungkinkan guru untuk melakukan diagnosis terhadap kelemahan siswa dan merancang perbaikan strategi pembelajaran yang lebih tepat sasaran. Mengingat peran vitalnya, sebuah instrumen evaluasi mutlak harus memenuhi persyaratan teknis yang ketat, meliputi aspek validitas untuk memastikan ketepatan pengukuran, reliabilitas untuk menjamin keajegan hasil, daya pembeda untuk mengklasifikasikan kemampuan siswa, serta tingkat kesukaran yang proporsional. Hanya dengan instrumen yang terstandarisasi inilah, hasil penilaian dapat benar-benar mencerminkan profil kemampuan peserta didik yang sesungguhnya secara akuntabel (Chasanah et al., 2025; Tibr et al., 2025).

Dalam konteks pendidikan vokasi atau Sekolah Menengah Kejuruan (SMK), tuntutan terhadap kualitas penilaian menjadi semakin spesifik dan mendesak karena orientasinya pada kesiapan kerja. Mata pelajaran produktif dan adaptif, seperti Ekonomi Bisnis dan Administrasi Umum, memegang peranan penting dalam membekali siswa dengan kompetensi dasar yang relevan dengan dinamika dunia industri. Karakteristik mata pelajaran ini menuntut penguasaan kemampuan yang tidak sederhana; siswa harus memiliki ketajaman analitis dan kecakapan pemecahan masalah atau *problem solving* yang aplikatif. Kurikulum di tingkat ini dirancang untuk menjembatani teori di kelas dengan praktik di lapangan kerja, sehingga instrumen penilaian yang digunakan tidak boleh hanya berhenti pada tataran pemahaman konsep dangkal. Evaluasi harus mampu memotret kesiapan siswa dalam menghadapi tantangan administratif dan ekonomis yang nyata, memastikan bahwa lulusan yang dihasilkan memiliki kualifikasi yang *link and match* dengan kebutuhan pasar tenaga kerja yang kompetitif dan terus berubah (Dolonseda & Sendiang, 2025; Harahap et al., 2025).

Sejalan dengan tuntutan kompetensi abad ke-21, paradigma pembelajaran di sekolah kejuruan kini sangat menekankan pada pengembangan kemampuan berpikir tingkat tinggi atau *Higher Order Thinking Skills* (HOTS) (Harahap et al., 2025; Mukhlisah et al., 2024; Putri et al., 2024). Pembelajaran yang berorientasi pada level ini menjadi instrumen vital untuk mencetak lulusan yang tidak hanya piawai dalam hal teknis, tetapi juga memiliki kemampuan berpikir kritis, kreatif, dan inovatif. Dalam situasi nyata di dunia kerja, kemampuan untuk menganalisis data, mengevaluasi argumen, dan menciptakan solusi baru jauh lebih berharga daripada sekadar kemampuan mengingat fakta. Oleh karena itu, instrumen penilaian pada pembelajaran berbasis kompetensi harus bertransformasi; dari sekadar menguji hafalan informasi atau *recall* memori, menjadi alat ukur kemampuan penalaran yang mendalam. Soal-soal yang diujikan harus mampu menstimulasi nalar siswa untuk menghubungkan berbagai konsep, mengambil keputusan yang tepat dalam situasi yang kompleks, serta memprediksi dampak dari sebuah keputusan ekonomi atau administrasi yang diambilnya (Astuti et al., 2024; Dhiefayanti & Mundir, 2025; Sinaga & Simbolon, 2025).

Untuk menjamin bahwa instrumen evaluasi benar-benar mampu mengukur kompetensi tingkat tinggi tersebut secara akurat, diperlukan sebuah mekanisme kontrol kualitas yang ketat melalui analisis statistik butir soal. Analisis butir soal merupakan sebuah prosedur sistematis yang digunakan untuk membedah karakteristik setiap *item* pertanyaan berdasarkan data empiris

hasil pengerjaan siswa. Proses ini sangat krusial karena sering kali soal yang dirancang oleh guru terlihat baik secara kualitatif, namun ternyata cacat secara kuantitatif ketika diujikan. Melalui analisis ini, dapat teridentifikasi apakah sebuah butir soal berfungsi dengan baik, apakah terlalu mudah atau terlalu sulit, dan apakah mampu membedakan siswa yang pandai dengan yang kurang pandai. Langkah ini memungkinkan guru untuk memutuskan nasib setiap butir soal: apakah layak dipertahankan, perlu direvisi redaksinya, atau harus dibuang sama sekali karena tidak valid. Tanpa proses ini, asesmen berisiko menjadi bias dan gagal memberikan gambaran objektif mengenai pencapaian belajar siswa.

Seiring dengan pesatnya kemajuan teknologi informasi, proses analisis butir soal yang dulunya rumit dan memakan waktu kini dapat dilakukan dengan lebih efisien dan akurat menggunakan bantuan perangkat lunak atau *software*. Salah satu alat bantu yang banyak digunakan dalam dunia pendidikan adalah aplikasi ANATES. Program komputer ini dirancang khusus untuk membantu para pendidik dalam mengolah data hasil tes dengan cepat, menyediakan informasi statistik yang sangat rinci dan komprehensif. Aplikasi ini mampu menyajikan data mengenai validitas butir, reliabilitas tes secara keseluruhan, efektivitas fungsi pengecoh atau *distractor*, daya pembeda, hingga tingkat kesukaran soal dalam hitungan detik. Keberadaan teknologi ini memberdayakan guru untuk melakukan revisi instrumen tes yang berbasis pada data empiris atau *data-driven*, bukan sekadar asumsi atau intuisi semata. Pemanfaatan teknologi analisis ini berkontribusi signifikan dalam meningkatkan mutu instrumen evaluasi, menjadikannya lebih terpercaya dan dapat dipertanggungjawabkan secara ilmiah (Aburizaizah, 2021; Arini et al., 2025; Nugroho et al., 2021).

Meskipun perangkat teknologi dan teori mengenai evaluasi yang baik telah tersedia, kenyataan di lapangan sering kali menunjukkan adanya kesenjangan yang lebar antara kondisi ideal dan praktik nyata. Idealnya, setiap butir soal yang diujikan kepada siswa telah melalui proses validasi dan analisis empiris yang ketat untuk menjamin kualitasnya. Namun, realitasnya, banyak guru di sekolah menengah kejuruan masih menghadapi kendala dalam menerapkan prosedur ini secara konsisten. Beban administratif yang tinggi dan keterbatasan waktu sering kali membuat penyusunan soal dilakukan secara instan, tanpa melalui tahap uji coba atau analisis mendalam. Banyak soal yang digunakan berulang-ulang dari tahun ke tahun tanpa evaluasi kinerja butir, atau sekadar mengambil dari buku teks tanpa penyesuaian konteks. Akibatnya, alat ukur yang digunakan sering kali tidak valid dalam mengukur kemampuan berpikir tingkat tinggi siswa, sehingga hasil evaluasi menjadi kurang akurat dalam menggambarkan kompetensi riil yang dimiliki peserta didik.

Berangkat dari urgensi untuk menjembatani kesenjangan tersebut, penelitian ini hadir dengan fokus utama untuk menganalisis secara mendalam kualitas butir soal pilihan ganda berbasis *Higher Order Thinking Skills* (HOTS). Objek spesifik penelitian ini adalah mata pelajaran Ekonomi Bisnis dan Administrasi Umum pada jenjang kelas XI, yang merupakan fase krusial dalam pematangan kompetensi kejuruan. Penelitian ini menawarkan nilai kebaruan dengan menerapkan analisis komprehensif yang mencakup empat pilar utama kualitas tes, yaitu validitas, reliabilitas, tingkat kesukaran, dan daya pembeda, dengan bantuan teknologi komputasi. Melalui kajian ini, diharapkan dapat diperoleh gambaran empiris mengenai profil kualitas soal yang digunakan di sekolah, sekaligus memberikan rekomendasi perbaikan yang konkret. Hasil penelitian ini diharapkan dapat menjadi rujukan praktis bagi para pendidik untuk meningkatkan kualitas instrumen evaluasi mereka, memastikan bahwa proses asesmen berjalan lebih objektif, efektif, dan benar-benar mendukung ketercapaian capaian pembelajaran yang diharapkan di Sekolah Menengah Kejuruan.

METODE PENELITIAN

Penelitian ini menerapkan pendekatan deskriptif kuantitatif sebagai kerangka kerja utama untuk mengevaluasi dan menganalisis kualitas empiris dari instrumen evaluasi pembelajaran yang dikembangkan. Fokus utama studi diarahkan secara spesifik pada analisis butir soal bentuk pilihan ganda yang berbasis *Higher Order Thinking Skills* (HOTS), khususnya pada mata pelajaran Ekonomi Bisnis dan Administrasi Umum untuk jenjang kelas XI. Pemilihan pendekatan ini didasarkan pada urgensi untuk mendapatkan gambaran yang objektif dan terukur mengenai karakteristik statistik dari setiap item soal yang diujikan. Adapun subjek penelitian yang dilibatkan dalam studi ini berjumlah 20 orang siswa kelas XI yang ditentukan sebagai responden uji coba. Para responden ini diberikan instrumen tes yang telah disusun secara sistematis mengacu pada indikator kompetensi dasar materi pembelajaran yang relevan. Data kuantitatif yang diperoleh dari respon siswa ini menjadi basis data utama yang akan diolah lebih lanjut untuk menentukan kelayakan instrumen sebagai alat ukur hasil belajar yang standar di lingkungan Sekolah Menengah Kejuruan.

Instrumen pengumpulan data yang digunakan dalam penelitian ini berupa tes objektif pilihan ganda yang dikembangkan dengan mengacu pada indikator kompetensi Kurikulum SMK yang berlaku saat ini. Untuk menjamin akurasi hasil evaluasi, seluruh data hasil tes siswa dianalisis menggunakan bantuan perangkat lunak *ANATES*. Proses analisis data difokuskan pada empat parameter psikometrik utama untuk menentukan kualitas butir soal, yakni validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. Pengujian validitas dilakukan dengan menggunakan teknik korelasi *item-total* guna memastikan keselarasan antara butir soal dengan konstruk kompetensi yang diukur. Sementara itu, reliabilitas instrumen diuji menggunakan teknik *internal consistency* untuk mengukur derajat keajekan atau stabilitas tes. Selanjutnya, tingkat kesukaran soal dihitung berdasarkan proporsi jawaban benar dari seluruh peserta tes, sedangkan analisis daya pembeda dilakukan untuk menilai kemampuan butir soal dalam membedakan kelompok siswa yang memiliki kemampuan kognitif tinggi dengan kelompok siswa yang berkemampuan rendah secara signifikan.

Hasil analisis statistik yang diperoleh kemudian dijadikan landasan pengambilan keputusan mengenai status setiap butir soal, apakah layak untuk dipertahankan, memerlukan revisi, atau harus dieliminasi karena tidak memenuhi standar. Hal ini sejalan dengan prinsip evaluasi pendidikan bahwa analisis butir bertujuan untuk menyeleksi soal yang valid dan membuang butir yang cacat fungsi. Secara operasional, prosedur penelitian ini dilaksanakan melalui lima tahapan sistematis yang berurutan. Tahap pertama dimulai dengan penyusunan kisi-kisi dan instrumen berdasarkan indikator kompetensi, dilanjutkan dengan pelaksanaan tes kepada responden terpilih di sekolah. Tahap ketiga meliputi pengolahan data hasil tes secara komputasi menggunakan *ANATES*, yang diteruskan dengan tahap interpretasi hasil analisis statistik terhadap karakteristik setiap butir soal. Tahap terakhir adalah penyusunan rekomendasi perbaikan instrumen evaluasi. Melalui rangkaian prosedur ini, kualitas instrumen dapat dinilai secara empiris guna mendukung peningkatan mutu asesmen berbasis *HOTS* di tingkat pendidikan vokasi.

HASIL DAN PEMBAHASAN

Penelitian ini menganalisis kualitas butir soal pilihan ganda berbasis HOTS pada mata pelajaran Ekonomi Bisnis dan Administrasi Umum kelas XI yang diujikan kepada 20 siswa. Instrumen terdiri dari 72 butir soal yang dianalisis menggunakan software *ANATES* untuk memperoleh data mengenai validitas, reliabilitas, daya pembeda, dan tingkat kesukaran.

1. Validitas

Validitas menunjukkan sejauh mana butir soal mampu mengukur kemampuan yang seharusnya diukur. Validitas dalam konteks analisis butir soal didasarkan pada korelasi item-total antara skor tiap butir dengan skor total tes. Uji validitas merupakan langkah penting dalam analisis butir soal untuk memastikan bahwa setiap pertanyaan dalam tes benar-benar mampu mengukur kompetensi yang ingin dicapai.

Tabel 1. Hasil Analisis Validitas

Kriteria Validitas	No. Butir Soal	Jumlah	Presentase
Sangat Tinggi	22, 24, 27, 30, 33, 36, 57	6	8,33%
Tinggi	15, 19, 25, 26, 35, 46, 47, 51, 52, 53, 54, 58, 68	12	19,44%
Cukup	3, 4, 7, 8, 9, 11, 12, 14, 18, 28, 29, 31, 32, 40, 41, 48, 50, 59, 62, 65, 69, 70	22	30,56%
Rendah	1, 2, 5, 6, 10, 13, 16, 17, 34, 38, 42, 43, 44, 45, 49, 60, 61, 63, 66, 67, 71, 72	22	33,33%
Sangat Rendah	20, 21, 23, 37, 39, 64, 55, 56	8	8,33%

Sumber: Data diolah (2025)

Berdasarkan data pada Tabel 1, sebagian besar butir berada pada kategori rendah (33,33%) dan cukup (30,56%), sedangkan kategori sangat tinggi dan sangat rendah masing-masing hanya 8,33%. Hasil ini menunjukkan bahwa masih banyak butir yang belum secara optimal mengukur kompetensi yang dituju dan memerlukan revisi untuk meningkatkan kesesuaian indikator instrumen.

2. Reliabilitas

Reliabilitas merupakan ukuran yang menyatakan tingkat kekonsistenan suatu soal. Dengan kata lain reliabilitas adalah sejauh mana butir soal mampu menghasilkan hasil yang sama pada uji coba secara berulang. Reliabilitas dihitung menggunakan koefisien internal consistency yang dihasilkan melalui ANATES.

Tabel 2. Hasil Analisis Reliabilitas

Indikator	Rata-Rata	Simpangan Baku	Korelasi Xy	Reliabilitas Tes
1	3,75	1,65	0,43	0,6
2	4,7	2,2	0,53	0,7
3	4,75	2,4	0,88	0,94
4	6,2	2,59	0,57	0,73
5	6,05	1,76	0,53	0,69
6	5,15	2,72	0,57	0,73
7	5,10	2,57	0,53	0,69
8	4,35	1,79	0,22	0,36

RATA-RATA	5,0125	2,115	0,5025	0,65
-----------	--------	-------	--------	------

Sumber: Data diolah (2025)

Berdasarkan tabel 2 nilai tertinggi terdapat pada indikator 3 sebesar 0,94 yang termasuk kategori sangat tinggi, sedangkan terendah pada indikator 7 sebesar 0,45 atau cukup. Secara umum, instrumen memiliki reliabilitas cukup hingga sangat tinggi, yang berarti instrumen relatif stabil, meskipun beberapa indikator masih perlu revisi struktur butir soal.

3. Daya Pembeda

Daya pembeda adalah kemampuan butir soal untuk membedakan siswa yang berkemampuan tinggi (memahami materi) dengan siswa yang berkemampuan rendah (kurang memahami materi). Perhitungan daya pembeda adalah pengukuran sejauh mana suatu butir soal mampu membedakan peserta didik yang sudah menguasai kompetensi dengan peserta didik yang belum/kurang menguasai kompetensi berdasarkan kriteria tertentu. Soal dengan daya pembeda baik akan memberikan gambaran objektif terhadap kemampuan peserta didik secara proporsional

Tabel 3. Hasil Analisis Daya Pembeda

Kategori	No. Butir Soal	Jumlah	Persentase
Sangat Baik	3, 15, 16, 22, 24, 25, 26, 27, 30, 33, 35, 36, 40, 46, 47, 51, 52, 53, 54, 57, 68, 61	22	30,56%
Baik	1, 2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 19, 28, 29, 31, 32, 37, 38, 39, 43, 44, 45, 48, 49, 50, 55, 56, 59, 62, 65, 66, 67, 69, 70, 71, 72, 63	39	54,17%
Cukup	18, 23, 34, 41, 42, 56, 58, 60	8	11,11%
Jelek	20, 21, 64	3	4,17%

Sumber: Data diolah (2025)

Berdasarkan tabel 3 mayoritas butir termasuk kategori baik (54,17%) dan sangat baik (30,56%), sehingga dapat disimpulkan bahwa sebagian besar soal sudah mampu membedakan peserta secara efektif. Namun terdapat 4,17% yang berkategori jelek dan harus diperbaiki atau direvisi ulang.

4. Tingkat Kesukaran

Tingkat kesukaran ditetapkan berdasarkan persentase peserta didik yang memberikan jawaban tepat. Semakin tinggi persentase peserta didik yang menjawab benar, maka semakin mudah butir soal tersebut dikategorikan. Sebaliknya, jika hanya sedikit peserta yang mampu menjawab dengan benar, soal tersebut digolongkan sebagai soal yang sukar. Informasi mengenai tingkat kesukaran sangat penting bagi pengembang instrumen karena membantu memastikan bahwa butir soal yang digunakan tidak terlalu mudah ataupun terlalu sulit, sehingga dapat mengukur kemampuan peserta didik secara lebih akurat dan proporsional. Analisis ini juga berperan dalam menjaga keseimbangan kualitas tes agar mampu membedakan kemampuan peserta secara optimal.

Tabel 4. Hasil Analisis Tingkat Kesukaran

Kriteria Tingkat Kesukaran	No. Butir Soal	Jumlah	Persentase

Sangat Mudah	20, 21, 28, 34, 50	5	6,94%
Mudah	10, 13, 29, 31, 40, 43	7	9,72%
Sedang	1, 2, 3, 4, 5, 6, 7, 8, 11, 14, 16, 17, 19, 22, 23, 27, 30, 32, 33, 35, 36, 37, 38, 39, 41, 42, 44, 45, 46, 47, 48, 49, 51, 52, 53, 54, 56, 57, 58, 59, 60, 61, 62, 63, 65, 67, 68, 69, 70	48	66,67%
Sukar	9, 12, 15, 18, 24, 25, 26, 64, 71, 72	12	16,67%

Sumber: Data diolah (2025)

Berdasarkan tabel 4 data menunjukkan bahwa kategori sedang mendominasi (66,67%), yang menunjukkan distribusi ideal dan seimbang. Soal berada dalam komposisi yang dapat mengukur kemampuan peserta secara proporsional.

5. Efektivitas Pengecoh

Efektivitas pengecoh berfungsi untuk mengetahui seberapa baik opsi jawaban yang salah dapat mengecoh siswa yang tidak menguasai materi.

Tabel 5. Hasil Analisis Efektivitas Pengecoh

Kriteria Kualitas Pengecoh	No. Butir Soal	Jumlah	Presentase
Sangat Baik	12, 38, 58, 62, 63, 32	6	8,33 %
Baik	2, 4, 10, 15, 33, 42, 46, 52, 55, 56, 59, 66	12	16,67%
Kurang Baik	6, 8, 16, 32, 35, 41, 43, 44, 45, 47, 53, 61, 64, 65, 68, 70, 71	17	23,61%
Buruk	1, 3, 5, 9, 13, 17, 26, 30, 36, 48, 49, 50, 51, 57, 67	15	20,83%
Sangat Buruk	7, 11, 14, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 31, 34, 37, 39, 40, 54, 60, 69	22	20,56%

Sumber: Data diolah (2025)

Berdasarkan tabel 5 analisis efektivitas pengecoh kategori sangat buruk dan buruk terlihat mendominasi yang apabila digabungkan mencapai 51,39% dari total soal. Sebanyak 22 butir soal (30,56%) tergolong dalam kriteria sangat buruk. Pada kategori ini tidak ada satupun opsi pengecoh yang berfungsi atau dipilih oleh siswa kelompok bawah yang ditandai dengan negatif. Kondisi ini mengindikasikan bahwa opsi jawaban salah yang disusun terlalu mencolok, tidak logis, atau tidak homogen dengan materi. Sehingga siswa dapat dengan mudah menebak kunci jawaban tanpa perlu memahami konsep secara mendalam. Selain itu, terdapat 15 butir soal dengan persentase 20,83% dengan kriteria buruk, dimana hanya satu pengecoh yang berfungsi efektif. Soal-soal pada kedua kategori ini sebaiknya direvisi total karena gagal menjalankan fungsi diskriminasi soal dengan baik.

Di sisi lain, butir soal yang memiliki kualitas pengecoh yang memadai (kategori Sangat Baik dan Baik) berjumlah 18 butir soal atau setara dengan 25% dari total keseluruhan. Rinciannya adalah 6 butir soal (8,33%) memiliki kualitas Sangat Baik (semua pengecoh

berfungsi) dan 12 butir soal (16,67%) ber kriteria Baik (tiga pengecoh berfungsi). Butir soal pada kelompok ini dinilai telah memenuhi standar konstruksi tes yang baik karena mampu menarik perhatian siswa yang tidak kompeten untuk memilih jawaban salah, sekaligus dapat diidentifikasi sebagai jawaban salah oleh siswa yang kompeten. Soal-soal ini layak dipertahankan dan dimasukkan ke dalam bank soal.

Selanjutnya, terdapat 17 butir soal (23,61%) yang masuk dalam kategori Kurang Baik. Pada soal jenis ini, setidaknya dua opsi pengecoh telah berfungsi, namun masih terdapat opsi lain yang tidak dipilih sama sekali oleh peserta tes. Terhadap butir soal ini, tindak lanjut yang disarankan adalah revisi minor, khususnya perbaikan redaksional pada opsi pengecoh yang tidak berfungsi agar menjadi lebih masuk akal dan meningkatkan daya kecohnya.

Pembahasan

Analisis mendalam terhadap validitas butir soal instrumen evaluasi mata pelajaran Ekonomi Bisnis dan Administrasi Umum menunjukkan adanya tantangan signifikan dalam konstruksi soal berbasis *Higher Order Thinking Skills* atau HOTS. Temuan data yang mengungkapkan bahwa dominasi butir soal berada pada kategori validitas rendah dan cukup mengindikasikan adanya ketidakselarasan antara item tes dengan indikator kompetensi yang hendak diukur. Dalam pengembangan instrumen berbasis kemampuan berpikir tingkat tinggi, validitas yang rendah sering kali disebabkan oleh ketidaktepatan operasional kata kerja atau ambiguitas dalam stimulus soal yang menyebabkan interpretasi ganda oleh peserta didik. Kondisi ini menyiratkan bahwa meskipun soal dirancang untuk menguji kemampuan analisis dan evaluasi, namun secara teknis konstruksi butir soal tersebut belum sepenuhnya mampu memotret kemampuan kognitif siswa yang sesungguhnya. Oleh karena itu, diperlukan tinjauan ulang terhadap kesesuaian antara kisi-kisi soal dengan redaksi pertanyaan agar instrumen benar-benar valid dalam mengukur variabel yang ditargetkan (MF et al., 2023; Ulya et al., 2022).

Meskipun terdapat catatan kritis pada aspek validitas, analisis reliabilitas instrumen menunjukkan konsistensi yang cukup memuaskan dengan kategori reliabilitas tes yang tergolong tinggi secara rata-rata. Hal ini menandakan bahwa instrumen tes tersebut memiliki tingkat keajegan atau stabilitas yang baik ketika digunakan secara berulang pada subjek yang sama. Namun, tingginya reliabilitas yang tidak dibarengi dengan validitas yang optimal pada seluruh butir soal menimbulkan paradoks evaluasi. Situasi ini dapat diartikan bahwa instrumen tersebut konsisten dalam menghasilkan skor, namun skor yang dihasilkan belum tentu akurat dalam merepresentasikan kompetensi HOTS siswa secara presisi. Kesenjangan antara reliabilitas dan validitas ini menjadi sinyal bagi pengembang soal untuk tidak hanya berfokus pada konsistensi internal tes, tetapi juga harus memastikan bahwa konten materi yang diujikan benar-benar relevan dan representatif terhadap kurikulum yang berlaku (Aryani et al., 2025; Yusuf, 2024).

Tinjauan terhadap tingkat kesukaran soal memperlihatkan distribusi yang ideal dengan dominasi soal pada kategori sedang yang mencapai lebih dari enam puluh persen. Proporsi ini sangat menguntungkan dalam konteks evaluasi pendidikan karena soal yang terlalu mudah tidak akan merangsang daya nalar siswa, sedangkan soal yang terlalu sukar dapat memicu frustrasi dan keputusan. Dalam kerangka *Higher Order Thinking Skills*, soal berkategori sedang harus dimaknai sebagai soal yang menuntut pemecahan masalah namun tetap terjangkau oleh nalar siswa kelas XI. Dominasi kategori ini menunjukkan bahwa instrumen tes telah dirancang dengan memperhatikan kemampuan rata-rata siswa, sehingga mampu memfasilitasi proses berpikir kritis tanpa membebani siswa dengan kompleksitas yang tidak perlu. Keseimbangan tingkat kesukaran ini menjadi modal penting untuk memetakan kemampuan siswa secara objektif tanpa bias kesulitan yang ekstrem (Masardi, 2025; Yuliastrin et al., 2023).

Salah satu kekuatan utama dari instrumen yang dikembangkan terletak pada daya pembeda soal yang sebagian besar berada pada kategori baik dan sangat baik. Temuan ini mengindikasikan bahwa butir-butir soal tersebut sangat efektif dalam mendiskriminasi atau membedakan antara kelompok siswa yang memiliki penguasaan materi tinggi dengan kelompok siswa yang penguasaannya rendah. Kemampuan diskriminasi yang tinggi sangat krusial dalam tes berbasis HOTS karena tujuan utama evaluasi adalah untuk mengidentifikasi level kognitif siswa secara akurat (Abdullah et al., 2025; Atana & Ansori, 2025; Saputra et al., 2025). Butir soal yang mampu membedakan kemampuan siswa membuktikan bahwa soal tersebut tidak bias dan memiliki parameter yang jelas dalam menentukan benar atau salahnya sebuah jawaban. Hal ini memberikan legitimasi bahwa skor yang diperoleh siswa benar-benar mencerminkan gradasi kemampuan mereka dalam memahami materi Ekonomi Bisnis dan Administrasi Umum (Muhson, 2020; Welsandt et al., 2024).

Namun, perhatian serius harus diarahkan pada efektivitas pengecoh atau *distractor* yang terbukti menjadi titik terlemah dalam instrumen ini. Data menunjukkan bahwa lebih dari setengah total butir soal memiliki kualitas pengecoh yang buruk dan sangat buruk. Kegagalan fungsi pengecoh ini berarti opsi jawaban salah yang disediakan tidak mampu menarik perhatian siswa yang kurang kompeten, sehingga mereka dapat dengan mudah menebak kunci jawaban yang benar. Dalam konstruksi soal pilihan ganda, *distractor* yang tidak berfungsi menjadikan tes kehilangan esensi kognitifnya karena peluang menebak menjadi sangat besar. Hal ini sangat kontraproduktif dengan prinsip HOTS yang menuntut analisis mendalam. Jika pengecoh terlalu implusif atau jelas salahnya, maka siswa tidak perlu menggunakan kemampuan berpikir tingkat tinggi untuk menemukan jawaban, melainkan cukup menggunakan metode eliminasi sederhana yang mereduksi kualitas evaluasi (Hikmawati, 2024; Salsabila et al., 2025; Widiarini et al., 2025).

Rendahnya kualitas pengecoh memiliki korelasi langsung dengan rendahnya validitas butir soal yang ditemukan pada analisis sebelumnya. Ketika opsi pengecoh tidak berfungsi secara logis atau homogen, maka struktur soal menjadi cacat secara psikometrik. Siswa yang seharusnya diuji kemampuan analisisnya menjadi terbantu oleh kelemahan konstruksi soal. Implikasi dari temuan ini menuntut adanya perbaikan radikal pada opsi-opsi jawaban yang salah. Pengecoh dalam soal HOTS seharusnya dirancang berdasarkan miskonsepsi umum yang sering dialami siswa atau logika yang tampak benar namun keliru secara konsep. Hanya dengan memperbaiki kualitas pengecoh agar lebih masuk akal dan menipu secara akademis, validitas soal dapat didongkrak naik. Tanpa perbaikan pada sektor ini, instrumen tes akan sulit untuk diandalkan sebagai alat ukur kemampuan berpikir kritis yang valid.

Secara keseluruhan, analisis butir soal menggunakan perangkat lunak *ANATES* ini menyimpulkan bahwa instrumen tes memiliki potensi yang baik dari aspek reliabilitas, tingkat kesukaran, dan daya pembeda, namun memerlukan revisi substansial pada aspek validitas dan efektivitas pengecoh. Instrumen ini layak untuk dipertahankan sebagian, terutama butir-butir yang memiliki daya pembeda tinggi, namun butir soal dengan pengecoh yang buruk harus direkonstruksi ulang atau diganti. Implikasi penelitian ini menegaskan bahwa menyusun soal HOTS tidak sekadar membuat pertanyaan sulit, tetapi membutuhkan teknik penyusunan opsi jawaban yang cermat dan terukur. Langkah perbaikan yang disarankan adalah melakukan revisi redaksional pada opsi jawaban dan melakukan uji coba ulang untuk memastikan bahwa setiap butir soal telah memenuhi standar psikometrik yang ketat sebelum digunakan dalam evaluasi sumatif yang sesungguhnya.

KESIMPULAN

Berdasarkan hasil analisis kualitas butir soal pilihan ganda berbasis HOTS pada mata pelajaran Ekonomi Bisnis dan Administrasi Umum kelas XI menggunakan software ANATES, dapat disimpulkan bahwa instrumen evaluasi yang digunakan memiliki kualitas yang cukup baik, namun masih memerlukan perbaikan pada aspek tertentu. Hasil analisis validitas menunjukkan bahwa sebagian besar butir berada pada kategori rendah (33,33%) dan cukup (30,56%), sehingga diperlukan peninjauan kembali aspek kesesuaian materi dan konstruksi soal. Pada aspek reliabilitas, instrumen berada pada kategori cukup hingga sangat tinggi, dengan nilai tertinggi pada indikator 3 sebesar 0,94 dan terendah pada indikator 7 sebesar 0,45, yang menunjukkan bahwa instrumen secara umum stabil tetapi beberapa indikator perlu diperbaiki. Hasil daya pembeda menunjukkan bahwa mayoritas soal berada dalam kategori baik (54,17%) dan sangat baik (30,56%), yang berarti sebagian besar butir mampu membedakan peserta berkemampuan tinggi dan rendah secara efektif. Namun terdapat 4,17% butir dalam kategori jelek yang harus direvisi. Pada aspek tingkat kesukaran, distribusi soal berada pada komposisi ideal dengan dominasi kategori sedang (66,67%), diikuti sukar (16,67%), mudah (9,72%), dan sangat mudah (6,94%). Secara keseluruhan, instrumen dapat dikategorikan layak, namun memerlukan evaluasi lanjutan untuk mencapai kualitas optimal.

DAFTAR PUSTAKA

- Abdullah, G., Isnanto, I., Marshanawiyah, A., & Ab-Rahman, M. S. (2025). Evaluasi pembelajaran IPA berbasis HOTS di SD Laboratorium UNG. *SCIENCE Jurnal Inovasi Pendidikan Matematika dan IPA*, 5(3), 1500. <https://doi.org/10.51878/science.v5i3.6927>
- Aburizaizah, S. J. (2021). Data-informed educational decision making to improve teaching and learning outcomes of EFL. *Journal of Education and Learning*, 10(5), 17. <https://doi.org/10.5539/jel.v10n5p17>
- Arini, A., Ratnawati, E., Nurlaili, N., Mulawarman, W. G., & Yahya, M. (2025). Penerapan total quality management (TQM) dalam meningkatkan mutu layanan pendidikan di sekolah: Studi kasus madrasah di Kalimantan Timur. *SOCIAL Jurnal Inovasi Pendidikan IPS*, 5(3), 1156. <https://doi.org/10.51878/social.v5i3.6937>
- Aryani, M., Abdullah, G., Nurainun, N., Sarlin, M., & Pomalingo, S. (2025). Pengaruh penggunaan aplikasi Google Earth terhadap motivasi belajar siswa pada mata pelajaran IPAS materi peta di kelas V SD Laboratorium UNG. *SOCIAL Jurnal Inovasi Pendidikan IPS*, 5(2), 744. <https://doi.org/10.51878/social.v5i2.6198>
- Astuti, N. N. P., Zubair, M., Alqadri, B., & Mustari, M. (2024). Internalisasi nilai-nilai dalam membentuk profil pelajar Pancasila di MA Manhalul Ma'arif Darek. *SOCIAL Jurnal Inovasi Pendidikan IPS*, 4(3), 155. <https://doi.org/10.51878/social.v4i3.3233>
- Atana, Y., & Ansori, I. (2025). Analisis implementasi Kurikulum Merdeka melalui model project based learning (PjBL) dalam pembelajaran IPAS kelas V SD Negeri 4 Gumiwang. *SCIENCE Jurnal Inovasi Pendidikan Matematika dan IPA*, 5(3), 1487. <https://doi.org/10.51878/science.v5i3.6929>
- Chasanah, U., Yatmi, Y., Hidayah, I., & Purwoko, P. (2025). Analisis instrumen assesment pembelajaran PAI berbasis multikultural. *LEARNING Jurnal Inovasi Penelitian Pendidikan dan Pembelajaran*, 5(3), 1413. <https://doi.org/10.51878/learning.v5i3.6653>
- Dhiefayanti, D. M., & Mundir, M. (2025). Analisis proses berfikir siswa SMP dalam menyelesaikan masalah statistika berdasarkan teori dual-process. *SCIENCE Jurnal*

- Inovasi Pendidikan Matematika dan IPA*, 5(4), 1658.
<https://doi.org/10.51878/science.v5i4.7228>
- Dolonseda, H. P., & Sendiang, D. J. (2025). Kajian empiris tentang kepemimpinan kepala sekolah, motivasi kerja, dan kinerja guru di SMK Negeri 1 Amurang. *SOCIAL Jurnal Inovasi Pendidikan IPS*, 5(4), 1482.
<https://doi.org/10.51878/social.v5i4.7969>
- Harahap, A. S., Siregar, N. S., Nasution, F. R. A., Yulastri, A., Ganefri, G., & Aditya, Y. (2025). Meta analisis pengaruh pendekatan edupreneurship pada pendidikan teknologi dan kejuruan. *LEARNING Jurnal Inovasi Penelitian Pendidikan dan Pembelajaran*, 5(3), 1040. <https://doi.org/10.51878/learning.v5i3.6625>
- Hikmawati, N. (2024). ASIAB chatbot for disaster mitigation learning at madrasah ibtidaiyah. *Geosfera Indonesia*, 9(2), 193. <https://doi.org/10.19184/geosi.v3i2.8126>
- Masardi, D. A. (2025). Penerapan model pembelajaran problem based learning berbantu media interaktif untuk meningkatkan keterampilan berpikir kritis dan hasil belajar IPAS peserta didik kelas 5 SDN Gogodalem 1. *SOCIAL Jurnal Inovasi Pendidikan IPS*, 5(3), 941. <https://doi.org/10.51878/social.v5i3.6865>
- Mf, A. F., Hala, Y., & Aziz, A. A. (2023). Uji kevalidan instrumen tes berorientasi higher order thinking skill (HOTS) pada materi sistem ekskresi kelas XI SMA/MA. *Jurnal Biogenerasi*, 8(1), 329. <https://doi.org/10.30605/biogenerasi.v8i1.2125>
- Muhson, A. (2020). Measuring student's economic behavior using situational judgment tests: Construct validity evidence. *Journal of Southwest Jiaotong University*, 55(6). <https://doi.org/10.35741/issn.0258-2724.55.6.29>
- Mukhlisah, I., Rochmawan, A. E., Aszahro, F., Wulandari, I. A. I., & Puspitaningrum, S. (2024). Pengembangan kreativitas dan ketrampilan motorik melalui media pembelajaran sistem tata surya di MI Sudirman Kadipiro Kecamatan Jumapolo Kabupaten Karanganyar. *COMMUNITY Jurnal Pengabdian Kepada Masyarakat*, 4(1), 56. <https://doi.org/10.51878/community.v4i1.3145>
- Nugroho, W., Yudha, R. P., Sundari, S., & Praja, H. N. (2021). Analisis instrumen asesmen unjuk kerja pada pembelajaran PJOK di sekolah dasar Kota Cirebon. *Gelanggang Olahraga Jurnal Pendidikan Jasmani dan Olahraga (JPJO)*, 4(2), 126. <https://doi.org/10.31539/jpjo.v4i2.1795>
- Putri, M. A. R. S., Herianto, E., Alqadri, B., & Sumardi, L. (2024). Pengaruh model pembelajaran PBL berbasis artificial intelligence terhadap hasil belajar siswa. *SOCIAL Jurnal Inovasi Pendidikan IPS*, 4(3), 145. <https://doi.org/10.51878/social.v4i3.3256>
- Salsabila, A., Ramadhani, C., & Faizin, M. S. (2025). Berpikir induktif sebagai dasar kompetensi sikap kritis bagi peserta didik generasi millennial abad 21. *CENDEKIA Jurnal Ilmu Pengetahuan*, 5(1), 264. <https://doi.org/10.51878/cendekia.v5i1.4465>
- Saputra, D., Meitriana, M. A., & Suadnyani, L. P. (2025). Upaya meningkatkan higher order thinking skills (HOTS) peserta didik melalui penerapan model pembelajaran problem based learning pada mata pelajaran IPS. *SOCIAL Jurnal Inovasi Pendidikan IPS*, 5(2), 816. <https://doi.org/10.51878/social.v5i2.6683>
- Sinaga, G. X., & Simbolon, E. (2025). Penerapan pembelajaran mendalam dalam meningkatkan keterampilan berpikir kritis siswa pada pelajaran Agama Katolik di Sekolah Menengah Negeri 1 Delitua. *LEARNING Jurnal Inovasi Penelitian Pendidikan dan Pembelajaran*, 5(3), 1192. <https://doi.org/10.51878/learning.v5i3.6127>
- Tibr, T. U., Fauzan, F., & Nurmaliyah, Y. (2025). Penerapan Kurikulum Merdeka pada pembelajaran Pendidikan Agama Islam dan Budi Pekerti di SMK Sahid Jakarta.

- LEARNING Jurnal Inovasi Penelitian Pendidikan dan Pembelajaran*, 5(3), 1442.
<https://doi.org/10.51878/learning.v5i3.6652>
- Ulya, H., Rahayu, R., Sa'dijah, C., & Qohar, A. (2022). Model asesmen kinerja pada pembelajaran etnomatematika berbasis proyek: Bagaimana kelayakannya? *AKSIOMA Jurnal Program Studi Pendidikan Matematika*, 11(2), 1019.
<https://doi.org/10.24127/ajpm.v11i2.4763>
- Welsandt, N. J., Fortunati, F., Winther, E., & Abs, H. J. (2024). Constructing and validating authentic assessments: The case of a new technology-based assessment of economic literacy. *Empirical Research in Vocational Education and Training*, 16(1). <https://doi.org/10.1186/s40461-024-00158-0>
- Widiarini, P., Rapi, N. K., Suastra, I. W., & Suma, K. (2025). Studi pendahuluan: Problematika pembelajaran fisika SMA. *SCIENCE Jurnal Inovasi Pendidikan Matematika dan IPA*, 5(1), 131. <https://doi.org/10.51878/science.v5i1.4430>
- Yuliastrin, A., Vebrianto, R., & Fiqri, M. (2023). The mapping students' critical thinking ability on vibration and wave material. *SEJ (Science Education Journal)*, 7(2), 89.
<https://doi.org/10.21070/sej.v7i2.1633>
- Yusuf, F. W. (2024). Analisis butir soal asesmen sumatif biologi materi perubahan lingkungan dengan menggunakan ANATES pada kelas X SMA. *LEARNING Jurnal Inovasi Penelitian Pendidikan dan Pembelajaran*, 4(1), 126.
<https://doi.org/10.51878/learning.v4i1.2776>