Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

Studi Kelayakan Kualitas Butir Latihan Soal HOTS Pilihan Ganda dalam Pembelajaran Praktikum Akuntansi Lembaga Pemerintahan di SMK Wijaya Putra Surabaya

Yobel Kriszaida Ebenheser[1], Sri Ayu Endang Setyoningrum[2], Revalina Aulia Sabrina[3], Vivi Pratiwi[4], Luqman Hakim[5]
[1,2,3,4,5]Program Studi Pendidikan Akuntansi, Universitas Negeri Surabaya
Surel: yobel.23044@mhs.unesa.ac.id

Abstract

This study aims to serve as an evaluation tool for the quality of multiple-choice practice questions based on *Higher Order Thinking Skills* (HOTS) in accounting practices at SMK Wijaya Putra Surabaya. The research employs both qualitative and quantitative descriptive methods. Anates software version 4.0 was utilized, with 15 multiple-choice questions analyzed. Data on validity, reliability, discriminatory power, difficulty level, and distractor effectiveness were collected from 30 students of classes XI-AKL-1 and XI-AKL-2 for evaluation. The results indicate that 33.33% of the questions have high validity, the reliability of the questions falls into the moderate category (0.50), 86.66% of the questions are of medium difficulty, and a significant portion of the distractors are categorized as poor. Consequently, this analysis identifies both the strengths and weaknesses of the assessment tool, contributing to the development of improved questions that align with educational standards.

Keyword: Item Analysis, HOTS, Validity, Reliability, Anates

Abstrak

Penelitian ini bertujuan sebagai alat evaluasi kualitas materi soal latihan opsi berganda berdasarkan *Higher Order Thinking Skill* (HOTS) dalam praktik akuntansi pada SMK Wijaya Putra Surabaya. Penelitian ini memakai deskripsi kualitatif dan kuantitatif. Perangkat lunak Anates versi 4.0. 15 butir soal digunakan untuk melakukan analisis soal latihan opsi berganda. Data validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan efektivitas pengecoh yang dikumpulkan dari 30 siswa kelas XI-AKL-1 dan XI-AKL-2 untuk dievaluasi. Hasilnya menunjukkan bahwa 33,33% soal memiliki validitas tinggi, reliabilitas soal berada dalam kategori cukup (0,50), 86,66% soal memiliki tingkat kesukaran sedang, dan besar bagian pengecoh masuk dalam kategori buruk. Hasilnya, analisis ini menemukan kelemahan dan kekuatan alat asesmen untuk membuat soal yang lebih baik dan sesuai dengan standar pembelajaran.

Kata Kunci: Analisis Butir Soal, HOTS, Validitas, Reliabilitas, Anates

PENDAHULUAN

Dengan pendidikan, manusia dapat mencapai aktualisasi diri dan mengembangkan semua potensi serta kemampuan yang dimilikinya dan lembaga pendidikan menjadi tempat berlangsungnya proses pembelajaran dan pendidikan tersebut (Baeng et al., 2022). Asesmen pembelajaran ialah elemen krusial dan utama dari alur pendidikan yang bertujuan mengevaluasi pencapaian kompetensi peserta didik serta menyediakan masukan yang bermanfaat terhadap pendidik dan para siswa. Instrumen asesmen yang baik tidak hanya berfungsi untuk mengukur hasil belajar tetapi juga mampu mendorong siswa meningkatkan skil berpikir tingkat tinggi biasa disebut HOTS (Higher Order Thinking Skills), sehingga mampu mengerti, mengaplikasikan,

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

menganalisis, serta mengevaluasi pengetahuan faktual, konseptual, prosedural, serta metakognitif berdasar tingkat perasaan keingin tahuan peserta didik tentang ilmu pengetahuan (Ansya, Alfianita, Syahkira, et al., 2024; Masrurah et al., 2020).

Magdalena et al (2021) menguraikan jenis tes tertulis objektif beberapa diantaranya terdiri dari soal pilihan ganda, isian, jawaban singkat, benar-salah, menjodohkan, dan uraian. Dalam penelitian ini, peneliti menggunakan tes pilihan ganda sebagai fokus penelitian karena tes ini dianggap lebih praktis dalam hal pelaksanaan dan penilaiannya, serta efektif dalam mengevaluasi hasil pembelajaran dengan terbatasnya waktu. Salah satu kelebihan tes pilihan ganda dibandingkan tes uraian adalah kemudahan dalam proses penilaiannya. Hal ini dikarenakan dalam tes pilihan ganda, jawaban yang benar dan salah sudah ditentukan sebelumnya, sehingga pemeriksaan dapat dilakukan dengan lebih jelas dan sederhana (Agustin et al., 2023). Selain itu, nilai pada tes pilihan ganda biasanya lebih konsisten dibandingkan dengan tes uraian, yang sering kali memiliki pembobotan berbeda untuk setiap soal (Ibrahim & Muslimah, 2021).

Meskipun tes pilihan ganda sering dianggap lebih mudah daripada tes uraian, tes ini memiliki tantangan dan kelemahannya sendiri. Salah satu tantangan selain soal dalam bentuk *Higher Order Thinking Skills* (HOTS) tersebut adalah guru harus menyusun opsi jawaban yang dapat mengecoh siswa (Purwati et al., 2021). Beberapa guru berpendapat adanya kelemahan pada pengukuran tes pilihan ganda pada kemampuan analitis dan keterampilan berpikir kritis siswa, sementara metode evaluasi lain, seperti tes esai, lebih efektif untuk menilai kemampuan siswa dalam mengorganisasi dan mengartikulasikan ide mereka secara mendalam (Agustin et al., 2023). Guru harus memastikan bahwa soal-soal tersebut disajikan dengan jelas dan tetap mampu menguji kemampuan berpikir siswa. Salah satu indikator keberhasilan proses pembelajaran pada peserta didik dapat dilihat melalui hasil belajar dan kegiatan evaluasi yang dilakukan oleh guru melalui tes tertulis maupun lisan dengan mempersiapkan seperangkat soal yang akan diberikan kepada peserta didik dengan kategori soal yang baik atau soal yang sudah di analisis baik dari tingkat kesukaran, daya pembeda dan efektivitas pengecoh (Ansya, Alfianita, & Syahkira, 2024; Rotama et al., 2020).

Oleh karena itu, penting untuk melakukan evaluasi kualitas instrumen asesmen guna memastikan bahwa soal yang digunakan relevan dan efektif dalam mendukung proses pembelajaran. Evaluasi penilaian yang terstruktur dapat membantu meningkatkan kualitas mutu asesmen, sehingga sesuai dengan tuntutan pendidikan modern dalam kurikulum merdeka. Berdasarkan pernyataan di atas diperoleh hasil akhir mengenai salah satu faktor penentu kemajuan suatu negara adalah sistem pengelolaan pendidikannya dan kualitas tenaga kerja yang dimilikinya. Untuk mendukung evaluasi tersebut, perangkat lunak seperti Anates versi 4.0 menjadi solusi yang efisien. Sari dan Herawati (2014), anates mampu menganalisis soal tes dengan berbagai cara, seperti: 1) Mengakumulasi perhitungan skor skor 2) mengukur reliabilitas tes 3) Menggolongkan subjek menjadi kelompok tinggi atau rendah 4) mengukur daya pembeda 5) menentukan nilai tingkat kesulitan 6) mengukur korelasi skor butir dengan skor total dan

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

7) menilai kualitas pengecoh (distraktor). Hasil dari analisis ini berguna untuk memperbaiki instrumen asesmen sehingga lebih sesuai dengan kebutuhan pembelajaran.

Penelitian ini bertujuan untuk mengevaluasi kualitas soal pilihan ganda berbasis HOTS dalam penilaian pembelajaran menggunakan Anates versi 4.0. Melalui analisis ini, diharapkan dapat ditemukan kelebihan dan kelemahan dari instrumen asesmen tersebut, sehingga dapat mendukung pengembangan pembelajaran yang lebih efektif dan membantu siswa mencapai kompetensi yang optimal. Selain itu, penelitian ini juga menyoroti pentingnya pembelajaran praktikum akuntansi untuk instansi pemerintah, yang merupakan topik yang jarang dibahas. Hasil penelitian diharapkan dapat menjadi referensi praktis bagi guru dan pendidik dalam merancang instrumen penilaian yang berkualitas, terutama dengan mengintegrasikan komponen HOTS ke dalam soal latihan. Hasil penelitian ini juga diharapkan dapat meningkatkan desain soal agar lebih relevan dengan kebutuhan mahasiswa akuntansi serta memperkaya pengetahuan mereka tentang akuntansi.

METODE PENELITIAN

Metode deskriptif kualitatif dan kuantitatif digunakan dalam penelitian ini. Metode penelitian deskriptif kualitatif adalah pendekatan penelitian yang bertujuan untuk memahami fenomena secara mendalam dengan menggambarkan atau mendeskripsikan data berdasarkan pandangan subjek penelitian (Sugiyono, 2013). Metode tersebut digunakan untuk mengevaluasi kualitas soal pilihan ganda akuntansi berbasis HOTS yang mampu menguji kualitas tiap butir soal dan analisis empiris butir soal menggunakan Anates Versi 4.0. Penelitian ini melibatkan siswa SMK Wijaya Putra Surabaya yang duduk di kelas XI sebagai subjek dengan jumlah siswa sebanyak 30 siswa dari kelas XI-AKL-1 dan XI-AKL-2. Penelitian pada bulan oktober 2024 ini terdiri dari lima belas soal pilihan ganda yang telah divalidasi oleh para ahli. Siswa mengisi soal-soal tersebut dengan platform Google Form sebagai metode pengumpulan data.

Proses penelitian dimulai dengan menentukan masalah, yang berarti diperlukan evaluasi terhadap kualitas soal berbasis HOTS. Selanjutnya, instrumen dibuat dan soal diberikan kepada siswa untuk pengumpulan data. Dengan menggunakan perangkat lunak Anates versi 4.0, data yang terkumpul dievaluasi validitas, reliabilitas, daya pembeda, tingkat kesukaran, dan efektivitas pengecoh. Hasil analisis digunakan untuk mengetahui kelemahan dan kelebihan soal, serta memberikan saran untuk pengembangan instrumen penilaian yang lebih baik.

HASIL DAN PEMBAHASAN

Hasil

Peneliti menyusun media evaluasi berbentuk soal pilihan ganda dan sudah divalidasi oleh dosen mata kuliah evaluasi belajar dan pembelajaran Vivi Pratiwi, S.Pd., M.Pd. Media penilaian tahap berikutnya ada 15 butir soal dikerjakan oleh 30 siswa/siswi kelas XI AKL 1 & XI AKL 2 di SMK Wijaya Putra Surabaya serta jawaban seluruh siswa dimasukkan ke dalam aplikasi Anates 4.0.9.

Hasil Uji validitas menggunakan aplikasi Anates melihat validitas butir soal melalui nilai korelasi antara skor

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

butir soal dengan skor keseluruhan, selain itu hasil analisis validitas dapat dilihat pada kolom keterangan tabel validitas Anates yang memberi tau signifikan atau tidak signifikan (tidak valid) dengan tanda (-) hasil analisis validitas soal nya. Setelah data diolah dengan Anates, dengan jumlah subyek 30, jumlah butir soal 15, serta bobot jawaban benar 2, diperoleh hasil data korelasi butir soal dengan skor total seperti pada tabel berikut.

Tabel 1. Hasil Analisis Uji Validitas

Rentang	Butir Soal	Jumlah	Presentase	Keterangan
0,400 - 0,600	9, 10, 12, 14, 15	5	33,33%	Cukup
0,200 - 0,400	2, 3, 4, 7, 11, 13	6	40%	Rendah
0,00 - 0,200	1, 5, 6, 8	4	26,66%	Sangat Rendah
Jumlah		15	100%	

Tabel 1 menunjukan adanya 5 butir soal cukup signifikan diantaranya butir soal no. 9, 10, 12, 14, dan 15, sisanya tidak signifikan. Maka dapat disimpulkan bahwa dari 15 soal terdapat 5 butir soal (33,33%) dengan kriteria valid dan 10 butir soal (66,66%) dengan kriteria tidak valid. Kevalidan soal menunjukkan bahwa kualitas soal yang dibuat mampu mencerminkan kesesuaian antara suatu pengukuran dan keefektifan tes tersebut. Butir soal dengan validitas tinggi menunjukkan tingkat keakuratan dalam mengukur kemampuan siswa. Sebaliknya, butir soal dengan validitas yang sangat rendah memerlukan tindak lanjut, karena soal tersebut dianggap tidak valid.

Berikut reliabilitas latihan soal mata pelajaran akuntansi keuangan lembaga dengan menggunakan aplikasi Anates versi 4.0.9.

Tabel 2. Nilai Reliabilitas

Rata-Rata	9,20
Simpangan Baku	2,35
Korelasi XY	0,33
Reliabilitas	0,50

Nilai reliabilitas menunjukkan besar nilai rata-rata, simpangan baku, korelasi XY, dan nilai reliabilitas tes 0,50. Tes dinilai handal apabila koefisien sekurang-kurangnya 0,80, sedangkan reliabilitas instrumen ini sebesar 0,50 maka tes ini termasu cukup reliabel sesuai dengan kriteria korelasi koefisiennya $0,40 \leq DB \leq 0,70$.

Data indeks daya pembeda yang diperoleh dari hasil analisis aplikasi Anates dinyatakan dalam tabel berikut.

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
https://doi.org/10.24114/jgk.v9i3.64381

DAYA PEMBEDA
==========

Jumlah Subyek= 30
Klp atas/bawah(n)= 8
Butir Soal= 15
Nama berkas: C:\USERS\YOBELK~1\DESKTOP\DATAKE~2.ANA

No Butir Baru	No Butir Asli	Kel. Atas	Kel. Bawah	Beda	Indeks DP (%)
1	1	7	4	3	37,50
2	2	8	5	3	37,50
3	3	6	4	2	25,00
4	4	5	2	3	37,50
5	5	7	5	2	25,00
6	6	6	5	1	12,50
7	7	7	4	3	37,50
8	8	4	3	1	12,50
9	9	7	2	5	62,50
10	10	6	3	3	37,50
11	11	7	4	3	37,50
12	12	6	1	5	62,50
13	13	7	3	4	50,00
14	14	7	2	5	62,50
15	15	7	3	4	50,00

Gambar 1. Hasil Daya Pembeda

Berdasar data dalam tabel tersebut menunjukkan variasi tingkat daya pembeda soal no. 9, 12, 13, 14, 15 dinyatakan kategori daya pembeda baik karena masuk rentang 0,40-0,70 dan soal no. 1, 2, 3, 4, 5, 7, 10, 11 dinyatakan kategori daya pembeda cukup karena masuk rentang 0,20-0,40 dan keduanya memenuhi kriteria sehingga perlu pembahasan lebih lanjut untuk dilakukan perbaikan. Sedangkan soal nomor 6 dan 8 dinyatakan kategori daya pembeda jelek sehingga harus diganti dengan soal yang lain. Metode yang digunakan untuk mengidentifikasi apakah suatu soal dapat membedakan siswa yang telah mencapai tujuan pembelajaran dengan yang belum dikenal dengan istilah daya pembeda (Hasan & Mukhlisa, 2023). Menurut Alpusari (2014), daya pembeda soal mencerminkan kemampuan soal untuk membedakan antara siswa yang berprestasi tinggi dan siswa yang berprestasi rendah. Dengan kata lain, daya pembeda berfungsi untuk mengidentifikasi siswa mana yang lebih mampu dibandingkan dengan siswa lainnya (Fiska et al., 2021).

Selanjutnya data tingkat kesukaran beserta interpretasinya disajikan di tabel berikut.

Tabel 3. Hasil Tingkat Kesukaran

No	Indeks	Kategori	Nomor Soal	Jumlah Soal	Presentase %
1	0,00 ≤ TK ≤ 0,30	Sukar	-	0	0
2	0,31 ≤ TK ≤ 0,70	Sedang	1, 2, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 15	13	86,66
3	0,71 ≤ TK ≤ 1,00	Mudah	5 & 7	2	13,33
	Total			**15**	**100**

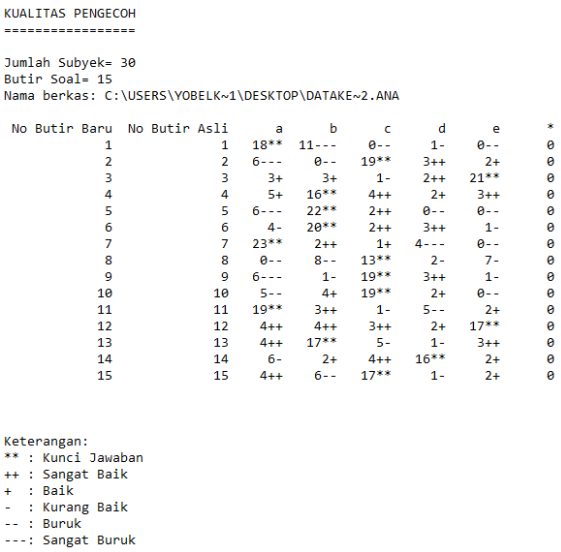
Dikaji dari tingkat kesukaran yang tertera pada tabel kita menjadi tahu jika tidak adanya latihan soal berkategori sukar, soal berkategori sedang ada 13 soal (86,66%) dan soal berkategori mudah ada 2 soal (13,33%). Maka, dapat disimpulkan bahwa semakin besar indeks, maka tingkat kesukaran yang diperoleh semakin mudah. Dengan demikian hasil penelitian menunjukkan bahwa tingkat kesukaran kategori sedang lebih dominan yakni sebesar 86,66%.

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

Dengan demikian proporsional tingkat kesukaran soal tersebut cukup seimbang.

Selanjutnya analisis kualitas pengecoh dengan aplikasi Anates dinyatakan dalam tabel berikut.



KUALITAS PENGECOH
==========

Jumlah Subyek= 30
Butir Soal= 15
Nama berkas: C:\USERS\YOBELK~1\DESKTOP\DATAKE~2.ANA

No Butir Baru	No Butir Asli	a	b	c	d	e	*
1	1	18**	11---	0--	1-	0--	0
2	2	6---	0--	19**	3++	2+	0
3	3	3+	3+	1-	2++	21**	0
4	4	5+	16**	4++	2+	3++	0
5	5	6---	22**	2++	0--	0--	0
6	6	4-	20**	2++	3++	1-	0
7	7	23**	2++	1+	4---	0--	0
8	8	0--	8--	13**	2-	7-	0
9	9	6---	1-	19**	3++	1-	0
10	10	5--	4+	19**	2+	0--	0
11	11	19**	3++	1-	5--	2+	0
12	12	4++	4++	3++	2+	17**	0
13	13	4++	17**	5-	1-	3++	0
14	14	6-	2+	4++	16**	2+	0
15	15	4++	6--	17**	1-	2+	0

Keterangan:
** : Kunci Jawaban
++ : Sangat Baik
+ : Baik
- : Kurang Baik
-- : Buruk
---: Sangat Buruk

Gambar 2. Hasil Kualitas Pengecoh

Berdasarkan hasil analisis latihan soal kualitas pengecoh, Pada soal 1 dan 6 banyak opsi pengecoh yang dinilai buruk (--) atau sangat buruk (---) karena hanya dipilih oleh sedikit atau bahkan tidak ada peserta. Sebaliknya, pada soal 11 dan 15 beberapa soal menunjukkan pengecoh yang efektif di mana opsi selain kunci memiliki daya tarik yang cukup kuat (+ atau ++). Dari data tersebut terlihat jelas adanya variasi kualitas soal ujian pada soal-soal ulangan harian. Sebagian besar soal, yaitu 8 soal (53%), masuk dalam kategori jelek, diikuti oleh 4 soal (27%) dikategorikan cukup, sedangkan 2 soal (13%) masuk dalam kategori sangat jelek, dan hanya 1 soal (7%) yang dianggap memiliki pengecoh yang baik.

Pada pertanyaan 1 dan 6, sebagian besar pemicu dinilai buruk atau sangat buruk karena hampir tidak ada peserta yang memilihnya. Di sisi lain, pertanyaan 11 dan 15 menunjukkan pengecoh yang cukup efektif karena beberapa pilihan selain kunci dapat menarik perhatian peserta yang tidak menguasai materi. Namun, soal 6 dan 8 menunjukkan masalah distribusi jawaban yang tidak merata, di mana sebagian besar peserta langsung memilih kunci jawaban tanpa terdistraksi oleh pengecoh. Hal ini mengindikasikan rendahnya daya pembeda soal, yang berarti perlu dilakukan revisi pengecoh agar soal lebih efektif dalam mengukur kemampuan siswa. Revisi ini dapat dilakukan dengan membuat pengecoh yang lebih relevan dan mirip dengan kunci jawaban, sehingga dapat menarik perhatian peserta yang kurang memahami materi.

Pembahasan

Berdasarkan hasil analisis validitas, diketahui bahwa hanya 5 dari 15 butir soal yang memenuhi kriteria validitas. Artinya, sebagian besar soal yang disusun tidak dapat dijadikan indikator yang tepat untuk mengukur kompetensi yang dituju dalam pembelajaran. Validitas merupakan

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

aspek penting dalam penyusunan instrumen, karena menunjukkan sejauh mana soal yang dibuat benar-benar sesuai dengan tujuan pembelajaran. Rendahnya validitas pada instrumen ini disebabkan oleh kurangnya relevansi antara soal dengan materi ajar serta tujuan kompetensi yang diharapkan. Beberapa soal kemungkinan masih bersifat mengulang hafalan atau tidak menuntut kemampuan berpikir tingkat tinggi (*Higher Order Thinking Skills*/HOTS). Kondisi ini menunjukkan perlunya revisi dan penyusunan soal yang lebih matang, agar setiap butir soal benar-benar mencerminkan indikator pencapaian kompetensi. Temuan ini sejalan dengan penelitian Elangovan & Sundaravel (2021) dan Suherman & Vidákovich (2022) yang menekankan bahwa validitas soal sangat penting untuk memastikan instrumen mampu mengukur apa yang seharusnya diukur dalam proses evaluasi pembelajaran.

Dari segi reliabilitas, instrumen memiliki koefisien sebesar 0,50. Nilai ini menunjukkan bahwa instrumen masih berada pada kategori cukup, artinya dapat digunakan namun belum memiliki konsistensi hasil yang tinggi. Reliabilitas yang rendah dapat menimbulkan perbedaan hasil yang signifikan jika soal digunakan berulang kali pada kelompok siswa yang berbeda, sehingga mengurangi keandalan data yang diperoleh. Hal ini menunjukkan bahwa soal perlu diperbaiki baik dari sisi konstruksi, redaksi, maupun variasi tingkat kesulitan agar hasil yang diperoleh lebih konsisten. Hasil ini sejalan dengan pendapat Ngo et al. (2024) dan Yan & Pastore (2022) yang menyatakan bahwa reliabilitas instrumen merupakan prasyarat utama untuk memastikan hasil evaluasi dapat dipercaya dan tidak dipengaruhi oleh faktor kebetulan semata.

Analisis tingkat kesukaran soal menunjukkan mayoritas butir berada pada kategori sedang, yakni sebesar 86,66%. Kondisi ini sebenarnya menunjukkan hal positif, karena soal berada pada tingkat kesukaran yang wajar, tidak terlalu mudah namun juga tidak terlalu sukar sehingga mampu menantang siswa tanpa menimbulkan beban berlebih. Soal pada kategori sedang biasanya efektif untuk membedakan siswa yang menguasai materi dengan yang belum. Namun demikian, masih ditemukan 2 soal yang tergolong mudah (13,33%), yang mengindikasikan adanya ketidakseimbangan variasi tingkat kesukaran. Idealnya, soal harus memiliki proporsi yang seimbang antara kategori mudah, sedang, dan sukar agar dapat menggambarkan distribusi kemampuan siswa secara lebih komprehensif. Temuan ini sejalan dengan penelitian(Alammary & Masoud (2025) dan Li et al. (2024), yang menyatakan bahwa komposisi tingkat kesukaran soal yang seimbang mampu meningkatkan daya ukur soal serta mendukung pencapaian tujuan pembelajaran berbasis HOTS.

Selain itu, hasil analisis menunjukkan bahwa efektivitas pengecoh pada sebagian besar soal masih rendah. Pengecoh yang kurang berfungsi menyebabkan pilihan jawaban yang salah tidak cukup menantang siswa yang tidak menguasai materi. Akibatnya, siswa dengan pengetahuan terbatas dapat dengan mudah menebak jawaban yang benar, sehingga mengurangi keakuratan instrumen dalam mengukur pemahaman siswa. Efektivitas pengecoh yang rendah menjadi salah satu kelemahan utama instrumen, karena pengecoh seharusnya

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

dirancang untuk mengantisipasi kesalahan konseptual yang umum dilakukan siswa. Hal ini sesuai dengan penelitian Martínez-Borreguero et al. (2024) dan Özdemir & Toker (2025), yang menjelaskan bahwa pengecoh yang baik dapat meningkatkan daya pembeda soal dan membantu pendidik mengetahui letak kesalahan konsep siswa dalam memahami materi.

Dengan demikian, hasil analisis validitas, reliabilitas, tingkat kesukaran, dan efektivitas pengecoh memberikan gambaran bahwa instrumen yang digunakan masih memerlukan banyak perbaikan. Revisi perlu difokuskan pada perbaikan redaksi soal, penyesuaian materi dengan kompetensi dasar, pengembangan variasi tingkat kesukaran soal, serta penyusunan pengecoh yang lebih berkualitas. Perbaikan ini penting dilakukan agar instrumen evaluasi benar-benar mampu mengukur ketercapaian pembelajaran, khususnya dalam kerangka implementasi kurikulum yang menekankan pengembangan kemampuan berpikir kritis dan HOTS. Hal ini sejalan dengan penelitian Magzoub et al. (2025) dan Puthiaparampil & Rahman (2021), yang menyatakan bahwa kualitas soal yang valid, reliabel, seimbang tingkat kesukarannya, dan memiliki pengecoh yang efektif sangat diperlukan untuk mendukung evaluasi pembelajaran yang komprehensif serta peningkatan mutu pendidikan.

KESIMPULAN

Hasil penelitian ini mengindikasikan bahwa meskipun terdapat beberapa soal HOTS yang telah memenuhi standar, secara keseluruhan kualitas soal tersebut masih jauh dari optimal. Validitas soal yang rendah menunjukkan bahwa banyak pertanyaan tidak dapat mengukur kemampuan berpikir tingkat tinggi siswa dengan efektif. Selain itu, reliabilitas instrumen yang berada dalam kategori sedang menandakan bahwa hasil penilaian dapat bervariasi secara signifikan antar siswa, yang dapat mempengaruhi keakuratan evaluasi kompetensi mereka.

Efektivitas pengecoh juga menjadi perhatian, karena beberapa pilihan jawaban tidak cukup menarik atau relevan untuk menguji pemahaman siswa secara mendalam. Oleh karena itu, disarankan untuk melakukan revisi menyeluruh terhadap soal-soal tersebut dengan memperhatikan aspek-aspek ini guna memastikan bahwa instrumen penilaian tidak hanya valid dan reliabel tetapi juga mampu mendorong siswa untuk berpikir kritis dan kreatif khususnya dalam konteks akuntansi keuangan lembaga/instansi pemerintah.

DAFTAR RUJUKAN

Agustin, R., Surani, D., Khasanah, A. N., Pratiwi, K. S., Nafizah, D., & Fajrin, R. I. M. (2023). PENGGUNAAN TES PILIHAN GANDA SEBAGAI ALAT EVALUASI DI SEKOLAH MENENGAH PERTAMA NEGERI 2 KEDAWUNG SRAGEN. *PANDU: Jurnal Pendidikan Anak Dan Pendidikan Umum*, *1*(4), 1–9.

Alammary, A., & Masoud, S. (2025). Towards Smarter Assessments: Enhancing Bloom's Taxonomy Classification with a Bayesian-Optimized Ensemble Model Using Deep Learning and TF-IDF Features. *Electronics*, *14*(12), 2312. https://doi.org/10.3390/electronics14122312

Alpusari, M. (2014). Analisis butir soal konsep dasar IPA 1 melalui

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

penggunaan program komputer anates versi 4.0 for Windows. *Primary*, *3*(2), 106–115.

Ansya, Y. A., Alfianita, A., & Syahkira, H. P. (2024). OPTIMIZING MATHEMATICS LEARNING IN FIFTH GRADES: THE CRITICAL ROLE OF EVALUATION IN IMPROVING STUDENT ACHIEVEMENT AND CHARACTER. *PROGRES PENDIDIKAN*, *5*(3), 302–311. https://prospek.unram.ac.id/index.php/PROSPEK/article/view/1120

Ansya, Y. A., Alfianita, A., Syahkira, H. P., & Syahrial, S. (2024). Peran Evaluasi Pembelajaran pada Mata Pelajaran Matematika Kelas V Sekolah Dasar. *Indiktika : Jurnal Inovasi Pendidikan Matematika*, *6*(2), 173–184. https://doi.org/10.31851/indiktika.v6i2.15030

Baeng, B., Situmorang, R., & Winarsih, M. (2022). Contextual electronics learning module in sociology learning at senior high school. *Journal of Education Research and Evaluation*, *6*(3), 509–519.

Elangovan, N., & Sundaravel, E. (2021). Method of preparing a document for survey instrument validation by experts. *MethodsX*, *8*, 101326. https://doi.org/10.1016/j.mex.2021.101326

Fiska, J. M., Hidayati, Y., Qomaria, N., & Hadi, W. P. (2021). Analisis butir soal ulangan harian IPA menggunakan software Anates pada pendekatan teori tes klasik. *Natural Science Education Research (NSER)*, *4*(1), 65–76.

Hasan, K., & Mukhlisa, N. (2023). Evaluation Program of Independent Curriculum in Elementary School: A New Curriculum in Indonesia. *Proceeding of The Progressive and Fun Education International Conference*, *8*(1), 61–69.

Ibrahim, I., & Muslimah, M. (2021). Tekhnik Pemeriksaan Jawaban, Pemberian Skor, Konversi Nilai dan Standar Penilaian. *Jurnal Al-Qiyam*, *2*(1), 1–9.

Li, D., Fan, X., & Meng, L. (2024). Development and validation of a higher-order thinking skills (HOTS) scale for major students in the interior design discipline for blended learning. *Scientific Reports*, *14*(1), 20287. https://doi.org/10.1038/s41598-024-70908-3

Magdalena, I., Mahromiyati, M., & Nurkamilah, S. (2021). Analisis instrumen tes sebagai alat evaluasi pada mata pelajaran sbdp siswa kelas ii sdn duri kosambi 06 pagi. *Nusantara*, *3*(2), 276–287.

Magzoub, M. E., Zafar, I., Munshi, F., & Shersad, F. (2025). Ten tips to harnessing generative AI for high-quality MCQS in medical education assessment. *Medical Education Online*, *30*(1). https://doi.org/10.1080/10872981.2025.2532682

Martínez-Borreguero, G., Naranjo-Correa, F. L., & Nuñez, M. M. (2024). Exploring color concepts in physics education: Addressing common preconceptions among teachers-in-training. *Color Research & Application*, *49*(3), 339–355. https://doi.org/10.1002/col.22919

Masrurah, M., Khaeruddin, K., & Bunga Dara, A. (2020). *Pengembangan Instrumen Asesmen Higher Order Thinking Skills (HOTS) pada Bidang Studi Fisika*.

Ngo, T. T. A., Bui, C. T., Chau, H. K. L., & Tran, N. P. N. (2024). Electronic

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025

Vol. 9 No. 3 Juni 2025, hlm 963-972
p-ISSN : 2548-883X ||e-ISSN : 2549-1288
https://jurnal.unimed.ac.id/2012/index.php/jgkp/article/view/64381
: https://doi.org/10.24114/jgk.v9i3.64381

word-of-mouth (eWOM) on social networking sites (SNS): Roles of information credibility in shaping online purchase intention. *Heliyon*, *10*(11), e32168. https://doi.org/10.1016/j.heliyon.2024.e32168

Özdemir, A. Z., & Toker, Z. (2025). Analysis of distractors in mathematics questions and their potential to lead misconceptions. *Thinking Skills and Creativity*, *56*, 101730. https://doi.org/10.1016/j.tsc.2024.101730

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 For Windows. *Jurnal Pendidikan UNIGA*, *15*(2), 460–473.

Puthiaparampil, T., & Rahman, M. (2021). How important is distractor efficiency for grading Best Answer Questions? *BMC Medical Education*, *21*(1), 29. https://doi.org/10.1186/s12909-020-02463-0

Rotama, A. D., Budiutomo, T. W., & Bowo, A. N. A. (2020). Analisis butir soal penilaian tengah semester mata pelajaran PPKN kelas VII di SMP Muhammadiyah 7 Yogyakarta. *Academy of Education Journal*, *11*(01), 24–35.

Sugiyono, S. (2013). *Metode penelitian pendidikan pendekatan kuantitatif, kualitatif, dan R&D*. Alfabeta.

Suherman, S., & Vidákovich, T. (2022). Assessment of mathematical creative thinking: A systematic review. *Thinking Skills and Creativity*, *44*, 101019. https://doi.org/10.1016/j.tsc.2022.101019

Yan, Z., & Pastore, S. (2022). Are teachers literate in formative assessment? The development and validation of the Teacher Formative Assessment Literacy Scale. *Studies in Educational Evaluation*, *74*, 101183. https://doi.org/10.1016/j.stueduc.2022.101183

Diterima pada : 16 Desember 2025 **Disetujui pada** : 25 Juni 2025 **Dipublikasi pada** : 30 Juni 2025