JURNAL PENDIDIKAN AKUNTANSI INDONESIA

Gabriel W.S.P., Galang P.S.R.,Mochammad R.A.S, Muhammad N., Vivi P., Luqman H.

Page 84 – 95

Vol . 23
No . 02
[2025]

ANALISIS KARAKTERISTIK BUTIR SOAL ELEMEN KOMPUTER AKUNTANSI FASE F MENGGUNAKAN SOFTWARE ANATES

ANALYSIS OF ITEM CHARACTERISTICS OF PHASE F ACCOUNTING COMPUTER ELEMENTS USING ANATES SOFTWARE

Gabriel Wahyu Satriyo Pidekso[1], Galang Pandu Satriya Ramadhan[2],Mochammad Ramadhan Adam Sampurno[3], Muhammad Nur Hafizh[4], Vivi Pratiwi[5], Luqman Hakim[6]
Universitas Negeri Surabaya[1,2,3,4,5,6]
24080304116@mhs.unesa.ac.id

Abstrak

Penilaian kualitas butir soal merupakan tahap krusial dalam memastikan efektivitas evaluasi pembelajaran, khususnya pada mata pelajaran elemen komputer akuntansi yang menuntut integrasi kompetensi teknis dan konseptual. Penelitian ini bertujuan menganalisis karakteristik butir soal fase F meliputi validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh dengan menggunakan software Anates pada tes pilihan ganda. Metode penelitian menggunakan pendekatan kuantitatif deskriptif dengan mengolah data hasil tes siswa berdasarkan teori tes klasik. Hasil penelitian menunjukkan bahwa validitas butir didominasi kategori cukup, reliabilitas instrumen sangat tinggi, dan tingkat kesukaran mayoritas berada pada kategori sedang. Daya pembeda sebagian besar berada pada kategori baik dan sangat baik, sedangkan efektivitas pengecoh menunjukkan variasi kualitas, dengan sebagian butir memerlukan revisi karena berada pada kategori kurang baik hingga sangat buruk. Temuan ini menegaskan bahwa instrumen memiliki kualitas yang baik secara keseluruhan, namun perlu perbaikan dalam validitas moderat dan pengecoh yang tidak berfungsi. Penelitian ini memberikan kontribusi penting terhadap pengembangan instrumen evaluasi akuntansi berbasis komputer yang lebih tepat, reliabel, dan diagnostik.

Kata kunci: Daya Pembeda; Pengecoh; Reliabilitas; Tingkat Kesukaran; Validitas

Abstract

Item quality evaluation is an essential step in ensuring the effectiveness of learning assessments, particularly in accounting computer elements that require the integration of technical and conceptual competencies. This study aims to analyze item characteristics in Phase F, including validity, reliability, difficulty level, discrimination power, and distractor effectiveness using the Anates software for multiple-choice tests. A descriptive quantitative approach was employed by processing student test results based on Classical Test Theory. The findings indicate that item validity was dominated by the moderate category, the test demonstrated very high reliability, and the difficulty level was largely categorized as moderate. Most items exhibited good to excellent discrimination power, while distractor effectiveness varied considerably, with several items requiring revision due to poor to very poor functioning distractors. These results confirm that although the instrument demonstrates strong psychometric properties overall, improvements are needed in moderately valid items and non-functioning distractors. This study contributes to the development of more accurate, reliable, and diagnostic assessment instruments for computer-based accounting education.

Key Words: *difficulty index, discrimination power, distractor effectiveness, reliability, validity*

PENDAHULUAN

Perkembangan teknologi informasi telah membawa transformasi besar dalam praktik akuntansi modern. Proses pencatatan, pelaporan, hingga analisis keuangan kini banyak ditopang oleh sistem akuntansi berbasis komputer yang menuntut penguasaan perangkat lunak akuntansi di berbagai jenjang pendidikan. Dalam konteks Sekolah Menengah Kejuruan (SMK), penguasaan kompetensi elemen komputer akuntansi menjadi bagian penting untuk memastikan kesiapan lulusan menghadapi kebutuhan industri yang semakin digital (Putri et al., 2024). Penelitian Sutrisno et al. (2023) menunjukkan bahwa pelatihan penggunaan software akuntansi untuk guru dan siswa SMK berhasil meningkatkan kompetensi dan profesionalisme, demikian juga Qurochman et al. (2024) menemukan bahwa pelatihan aplikasi komputer akuntansi di SMK mampu mempersiapkan siswa menghadapi tuntutan dunia kerja. Oleh karena itu, keberhasilan pembelajaran tidak hanya ditentukan oleh penguasaan teori akuntansi, tetapi juga oleh kemampuan peserta didik dalam mengaplikasikan prinsip akuntansi melalui teknologi secara efektif dan efisien.

Salah satu faktor penting dalam menjamin keberhasilan pembelajaran adalah evaluasi hasil belajar melalui tes dengan butir soal yang berkualitas. Yusuf (2024) menjelaskan bahwa kualitas tes dapat dilihat dari validitas, reliabilitas, tingkat kesukaran, daya pembeda, serta efektivitas pengecoh yang menentukan sejauh mana butir soal mampu mengukur kemampuan peserta didik secara objektif. Analisis karakteristik butir soal dengan bantuan perangkat lunak seperti Anates memudahkan guru dan peneliti dalam menilai kualitas instrumen evaluasi pembelajaran secara empiris, termasuk ukuran validitas, reliabilitas, tingkat kesukaran, dan daya pembeda (Fiska et al., 2021; Sabela et al., 2025). Hasil penelitian Hermaya et al. (2024) dan Mawardi et al. (2023) membuktikan bahwa penggunaan Anates memberikan hasil analisis yang akurat, membantu mengidentifikasi butir yang tidak sesuai standar, serta memperkuat validitas instrumen evaluasi pembelajaran. Temuan-temuan ini menegaskan pentingnya pendekatan berbasis data empiris untuk meningkatkan mutu penilaian hasil belajar, terutama pada bidang kejuruan.

Berbagai studi terdahulu telah membahas penerapan Anates pada berbagai mata pelajaran. Ahmad et al. (2024) menemukan bahwa sebagian besar butir soal berbasis Higher Order Thinking Skills (HOTS) pada mata pelajaran komputer akuntansi di SMKN 10 Surabaya berada pada kategori baik, walaupun masih terdapat butir dengan daya pembeda rendah. Penelitian Agnola et al. (2025) di SMK Ketintang juga menunjukkan bahwa sebagian besar soal akuntansi komputer memiliki tingkat kesukaran dan reliabilitas yang sesuai, tetapi efektivitas pengecoh perlu diperbaiki. Kajian serupa di bidang sains oleh Yusuf (2024) dan Sheptian et al. (2025) menunjukkan bahwa sebagian besar butir memiliki validitas tinggi namun masih lemah pada aspek pengecoh. Dari sisi pendekatan metodologis, penelitian Wilsa et al. (2023) dan Mulyani et al. (2020) mengungkapkan bahwa sebagian besar studi masih menggunakan teori tes klasik tanpa mengombinasikannya dengan pendekatan modern. Berdasarkan kajian tersebut, dapat disimpulkan bahwa penelitian tentang analisis butir soal menggunakan Anates telah banyak dilakukan, namun masih berfokus pada mata pelajaran umum dan belum secara mendalam meneliti konteks kejuruan akuntansi.

Kesenjangan penelitian yang muncul dapat dilihat dari tiga dimensi. Pertama, secara konseptual, kajian sebelumnya belum banyak menyoroti karakteristik butir soal pada fase F elemen komputer akuntansi, yaitu fase yang menuntut integrasi kemampuan teknis penggunaan software akuntansi dengan pemahaman konsep akuntansi keuangan (Agnola et al., 2025). Kedua, secara empiris, penelitian yang menelaah butir soal akuntansi komputer masih terbatas jumlahnya dibandingkan bidang lain seperti IPA dan matematika (Hermaya et al., 2024; Mardiyyaningsih et al., 2023). Ketiga, secara metodologis, analisis yang dilakukan pada penelitian sebelumnya belum sepenuhnya menilai efektivitas pengecoh dan daya pembeda secara komprehensif, padahal kedua aspek tersebut penting untuk menilai kemampuan soal dalam membedakan siswa berkemampuan tinggi dan rendah (Alemu et al., 2024). Dengan demikian, penelitian ini hadir untuk menanggapi kesenjangan tersebut melalui analisis karakteristik butir soal elemen komputer akuntansi fase F dengan menggunakan software Anates secara sistematis dan mendalam.

Tujuan utama penelitian ini adalah untuk menganalisis kualitas butir soal elemen komputer akuntansi fase F yang meliputi aspek validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh dengan memanfaatkan software Anates. Penelitian ini berkontribusi pada pengayaan kajian evaluasi pembelajaran kejuruan, khususnya dalam konteks akuntansi berbasis teknologi. Secara praktis, hasil penelitian diharapkan dapat menjadi acuan bagi guru dan pengembang kurikulum SMK dalam menyusun serta merevisi butir soal yang lebih valid, reliabel, dan diagnostik. Dengan demikian, penelitian ini tidak hanya meningkatkan mutu asesmen pembelajaran akuntansi, tetapi juga mendukung penguatan sistem evaluasi pendidikan kejuruan di era digital.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode deskriptif. Pendekatan kuantitatif digunakan karena penelitian ini berfokus pada analisis data numerik hasil tes untuk menilai karakteristik butir soal secara empiris, sedangkan metode deskriptif digunakan untuk menggambarkan kualitas butir soal berdasarkan hasil analisis yang diperoleh dari perangkat lunak Anates. Menurut Sugiyono (2022), penelitian kuantitatif deskriptif bertujuan untuk mendeskripsikan suatu fenomena secara sistematis dan faktual berdasarkan data yang dapat diukur. Dengan demikian, penelitian ini bertujuan memberikan gambaran objektif mengenai validitas, reliabilitas, tingkat kesukaran, daya pembeda, serta efektivitas pengecoh pada butir soal elemen komputer akuntansi fase F. Subjek dalam penelitian ini adalah butir soal elemen komputer akuntansi fase F yang digunakan oleh peserta didik pada jenjang Sekolah Menengah Kejuruan (SMK). Data penelitian diperoleh dari hasil tes peserta didik yang telah dikerjakan secara keseluruhan. Instrumen penelitian berupa tes pilihan ganda yang terdiri atas sejumlah butir soal yang dikembangkan oleh guru mata pelajaran komputer akuntansi. Data tersebut kemudian diolah menggunakan software Anates versi terbaru, yang berfungsi untuk menganalisis karakteristik setiap butir soal meliputi tingkat kesukaran, daya pembeda, validitas, reliabilitas, serta efektivitas pengecoh. Analisis dilakukan berdasarkan teori tes klasik (Classical Test Theory) sebagaimana digunakan dalam penelitian sebelumnya (Fiska et al., 2021; Mawardi et al., 2023).

JURNAL PENDIDIKAN AKUNTANSI INDONESIA
Gabriel W.S.P., Galang P.S.R.,Mochammad R.A.S, Muhammad N., Vivi P., Luqman H.
Page 84 – 95
Vol . 23
No . 02
[2025]

Prosedur penelitian meliputi tiga tahap utama, yaitu pengumpulan data, analisis data, dan interpretasi hasil. Tahap pengumpulan data dilakukan dengan menghimpun lembar jawaban siswa serta kunci jawaban tes. Tahap analisis data dilakukan secara kuantitatif dengan bantuan program Anates untuk memperoleh nilai indeks setiap indikator karakteristik butir soal. Hasil analisis kemudian diinterpretasikan berdasarkan kriteria yang telah ditetapkan oleh Arikunto (2019), di mana kategori validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh diklasifikasikan ke dalam kategori sangat baik, baik, cukup, kurang, atau jelek. Hasil akhir dari analisis ini disajikan secara deskriptif untuk memberikan rekomendasi terhadap butir soal yang perlu direvisi, dipertahankan, atau dibuang dalam rangka meningkatkan kualitas evaluasi pembelajaran akuntansi berbasis komputer.

HASIL PENELITIAN DAN PEMBAHASAN

Hasil Penelitian

Hasil analisis validitas butir pada instrumen tes yang tersaji pada tabel 1 menunjukkan adanya variasi kualitas yang mencerminkan sejauh mana setiap butir mampu mengukur konstruk yang dimaksud secara tepat. Pada kategori *sangat tinggi*, terdapat 2 butir soal (5,56%), yang menandakan bahwa hanya sebagian kecil item memiliki korelasi sangat kuat dengan skor total. Meskipun jumlahnya kecil, keberadaan butir dengan validitas sangat tinggi menjadi indikator bahwa sebagian konstruk telah terwakili dengan sangat baik. Selanjutnya, kategori *tinggi* mencakup 9 butir soal (25%), yang menunjukkan bahwa butir-butir tersebut sudah memenuhi standar kelayakan psikometris dan dapat berkontribusi secara signifikan terhadap keakuratan pengukuran. Tingkat validitas tinggi sangat penting untuk memastikan instrumen dapat memberikan interpretasi hasil yang sahih (Levacher et al., 2023).

Tabel 1. Hasil Analisis Validitas

Kriteria Validitas	No. Butir Soal	Jumlah	Persentase
Sangat Tinggi	31, 34	2	5,56%
Tinggi	1, 14, 17, 24, 25, 28, 30, 33, 35	9	25%
Cukup	6, 7, 8, 9, 10, 11, 15, 16, 18, 19, 20, 22, 23, 26, 29, 32, 36	17	47,22%
Rendah	3, 12	2	5,56%
Sangat Rendah	2, 4, 5, 13, 21, 27	6	16,67%

Sumber: Data diolah oleh peneliti (2025)

Kategori *cukup* mendominasi distribusi dengan 17 butir soal (47,22%), menunjukkan bahwa sebagian besar item memiliki validitas moderat. Meski masih dapat digunakan, butir-butir ini idealnya mendapat revisi untuk menguatkan hubungan antara skor butir dan skor total, terutama jika bertujuan untuk asesmen berskala tinggi (Kreitchmann et al., 2024). Sementara itu, kategori *rendah* mencakup 2 butir soal (5,56%), mengindikasikan bahwa item tersebut kurang mampu merepresentasikan kompetensi yang diukur. Terakhir, kategori *sangat rendah* terdiri dari 6 butir soal

(16,67%), yang menunjukkan bahwa sejumlah item memiliki kontribusi minimal dan bahkan berpotensi menurunkan validitas keseluruhan tes. Butir dengan korelasi sangat rendah sebaiknya direvisi secara substansial atau dieliminasi (Toma et al., 2024).Secara keseluruhan, hasil ini menegaskan perlunya perbaikan pada sebagian butir untuk meningkatkan kesesuaian konstruk dan memperkuat kualitas instrumen secara menyeluruh.

Hasil analisis reliabilitas tes yang tersaji pada tabel 2 menunjukkan bahwa instrumen memiliki kualitas pengukuran yang sangat baik. Nilai rata-rata skor sebesar 23,85 menggambarkan kecenderungan pencapaian peserta yang berada pada tingkat sedang, sehingga instrumen memiliki rentang informasi yang cukup merata di seluruh tingkat kemampuan. Nilai simpang baku sebesar 7,26 menunjukkan adanya variasi skor yang cukup luas antar peserta, yang secara psikometris merupakan kondisi ideal untuk memastikan bahwa tes mampu menangkap perbedaan kemampuan secara nyata (Jahrami et al., 2024). Variasi skor yang memadai berperan penting dalam meningkatkan ketepatan estimasi reliabilitas karena memperbesar proporsi varians sejati dibandingkan varians kesalahan.

Tabel 2. Hasil Analisis Reliabilitas

Keterangan	Nilai
Rata-rata	23,85
Simpang Baku	7, 26
Korelasi XY	0,88
Reliabilitas Tes	0,93

Sumber: Data diolah oleh peneliti (2025)

Nilai korelasi XY sebesar 0,88 mengindikasikan hubungan sangat kuat antara skor butir dengan skor total, yang berarti sebagian besar item memberikan kontribusi signifikan terhadap konstruk yang diukur. Korelasi yang tinggi pada tahap analisis butir merupakan indikator konsistensi internal yang baik dan menunjukkan kesesuaian item dengan tujuan pengukuran (Marianti, 2023). Selanjutnya, nilai reliabilitas tes sebesar 0,93 mengonfirmasi bahwa instrumen berada pada kategori *sangat reliabel*. Dalam konteks asesmen pendidikan, reliabilitas di atas 0,90 sering diinterpretasikan sebagai bukti kuat bahwa tes memberikan hasil yang stabil, konsisten, dan minim kesalahan pengukuran (A. Stephen Editor, 2024). Secara keseluruhan, hasil tersebut menegaskan bahwa instrumen tes memiliki kualitas psikometris yang tinggi. Dengan reliabilitas sangat baik, variasi skor yang optimal, dan kontribusi item yang kuat terhadap skor total, instrumen ini layak digunakan untuk evaluasi pembelajaran yang membutuhkan akurasi dan konsistensi tinggi.

Hasil analisis tingkat kesukaran yang tersaji pada tabel 3 menunjukkan distribusi karakteristik butir yang cukup bervariasi, menggambarkan sejauh mana setiap item mampu memberikan tantangan yang proporsional kepada peserta tes. Pada kategori *sangat sukar*, hanya terdapat 2 butir soal (5,56%), sedangkan kategori *sukar* mencakup 1 butir soal (2,78%). Proporsi yang rendah pada kedua kategori ini menandakan bahwa sebagian besar butir tidak memberikan tingkat kesulitan ekstrem. Dalam evaluasi pendidikan yang berkualitas, jumlah butir terlalu sukar memang tidak disarankan karena dapat menurunkan motivasi dan tidak selalu mencerminkan kompetensi sebenarnya (Kim et al., 2024)

Tabel 3. Hasil Analisis Tingkat Kesukaran

Kriteria Tingkat Kesukaran	No. Butir Soal	Jumlah	Persentase
Sangat Sukar	2, 27	2	5,56%
Sukar	13	1	2,78%
Sedang	1, 3, 4, 5, 6, 8, 21, 24, 25, 28, 29, 31, 32, 33, 34, 35, 36	17	47,22%
Mudah	7, 9, 11, 14, 15, 16, 17, 23, 30	9	25%
Sangat Mudah	10, 12, 18, 19, 20, 22, 26	7	19,44%

Sumber: Data diolah oleh peneliti (2025)

Kategori *sedang* mendominasi distribusi dengan 17 butir soal (47,22%), menunjukkan bahwa sebagian besar item berada pada tingkat kesukaran ideal. Butir pada tingkat sedang dianggap paling efektif dalam mengukur kemampuan peserta secara proporsional karena memberikan peluang yang seimbang bagi peserta berkemampuan tinggi maupun rendah untuk menunjukkan performanya (Zhang & Colvin, 2024). Dominasi pada kategori ini mencerminkan kualitas konstruksi tes yang baik. Kategori *mudah* mencakup 9 butir soal (25%), yang menunjukkan adanya proporsi item dengan kesukaran rendah yang masih dapat diterima, terutama jika tes ditujukan untuk mengevaluasi penguasaan dasar. Namun, jumlahnya perlu tetap diawasi agar tidak menurunkan daya beda keseluruhan instrumen. Sementara itu, kategori *sangat mudah* mencakup 7 butir soal (19,44%), menunjukkan bahwa beberapa item terlalu sederhana sehingga tidak mampu menantang peserta untuk menunjukkan kemampuan optimal. Butir sangat mudah berpotensi menurunkan reliabilitas tes karena minimnya variasi respons peserta (Rezigalla et al., 2024). Secara keseluruhan, hasil analisis menunjukkan bahwa sebagian besar butir soal memiliki tingkat kesukaran yang baik, meskipun sejumlah item pada kategori ekstrem membutuhkan revisi untuk meningkatkan kualitas instrumen secara keseluruhan.

Hasil analisis daya pembeda yang tersaji pada tabel 4 menunjukkan distribusi kualitas butir soal yang beragam, mencerminkan sejauh mana setiap item mampu membedakan peserta didik berkemampuan tinggi dan rendah. Pada kategori *sangat baik*, terdapat 11 butir soal (30,56%), yang menunjukkan bahwa butir-butir tersebut memiliki kemampuan optimal dalam memisahkan kelompok peserta dengan pemahaman materi yang kuat dari mereka yang kurang menguasai. Daya pembeda yang tinggi menjadi indikator penting keefektifan instrumen tes, karena secara langsung berkaitan dengan validitas konstruk dan ketepatan interpretasi hasil asesmen (Awalurahman & Budi, 2024).

Tabel 4. Hasil Analisis Daya Pembeda

Kriteria Daya Pembeda Soal	No. Butir Soal	Jumlah	Persentase
Sangat Baik	9, 14, 24, 25, 28, 29, 30, 31, 33, 34, 35	11	30,56%
Baik	1, 3, 5, 6, 7, 11, 12, 15, 16, 17, 18, 23, 26, 32, 36	15	41,67%
Cukup	2, 4, 8, 10, 19, 20, 22	7	19,44%
Negatif	13, 21, 27	3	8,33%

Sumber: Data diolah oleh peneliti (2025)

Kategori *baik* merupakan kelompok terbesar dengan 15 butir soal (41,67%), menandakan bahwa sebagian besar item telah memenuhi standar minimal sebagai butir yang mampu memfasilitasi pembedaan kemampuan peserta secara memadai. Butir dalam kategori ini tetap layak digunakan tanpa revisi besar, meskipun beberapa perbaikan minor dapat meningkatkan performanya lebih lanjut (Nurjanah et al., 2024). Pada kategori *cukup*, terdapat 7 butir soal (19,44%), yang mencerminkan bahwa daya pembeda butir tersebut masih berada pada level moderat. Item pada kategori ini sebaiknya direvisi untuk memperjelas indikator kompetensi atau meningkatkan kualitas pengecoh, agar lebih efektif dalam memisahkan kemampuan peserta (Hakim, 2023). Sementara itu, kategori *negatif* mencakup 3 butir soal (8,33%), menunjukkan bahwa butir tersebut gagal berfungsi sebagaimana mestinya karena peserta berkemampuan rendah cenderung menjawab benar lebih sering daripada peserta berkemampuan tinggi. Kondisi ini merupakan indikasi kuat bahwa butir harus direvisi secara menyeluruh atau bahkan dieliminasi (Kumar et al., 2021).Secara keseluruhan, temuan ini menggambarkan bahwa sebagian besar butir telah memenuhi standar kualitas yang baik, namun sejumlah butir memerlukan perbaikan agar instrumen tes menjadi lebih valid dan reliabel.

Hasil analisis efektivitas pengecoh yang tersaji pada tabel 5 menunjukkan variasi kualitas yang signifikan pada instrumen tes yang diukur. Berdasarkan temuan, hanya 4 butir soal (11,11%) yang berada pada kategori *sangat baik*, menandakan bahwa sebagian kecil pengecoh berfungsi optimal dalam menarik perhatian peserta didik dari kunci jawaban. Pengecoh yang berfungsi dengan baik mampu mengalihkan respons peserta, terutama yang memiliki kemampuan rendah, sehingga meningkatkan daya pembeda butir (Rezigalla et al., 2024). Selanjutnya, terdapat 8 butir soal (22,22%) pada kategori *baik*, yang menunjukkan bahwa pengecoh pada kategori ini masih bekerja sesuai fungsi meskipun tidak seefektif kategori tertinggi.

Tabel 5. Hasil Analisis Efektivitas Pengecoh

Kriteria Efektivitas Pengecoh	No. Butir Soal	Jumlah	Persentase
Sangat Baik	9, 28, 32, 33	4	11,11%
Baik	3, 17, 23, 25, 29, 30, 31, 35	8	22,22%
Kurang baik	1, 4, 5, 7, 14, 16, 24, 27, 34, 36	10	27,78%
Buruk	2, 6, 8, 11, 13, 21	6	16,67%
Sangat Buruk	10, 12, 15, 18, 19, 20, 22, 26	8	22,22%

Sumber: Data diolah oleh peneliti (2025)

Kategori *kurang baik* merupakan kelompok terbesar dengan 10 butir soal (27,78%), mengindikasikan masih adanya kelemahan dalam penyusunan pengecoh. Pengecoh yang tidak dipilih secara proporsional sering muncul akibat redaksi yang tidak menarik, kurang relevan, atau terlalu mudah ditebak (Lee et al., 2025). Selain itu, 6 butir soal (16,67%) termasuk kategori *buruk*, yang menunjukkan bahwa sebagian besar pengecoh pada butir tersebut tidak berfungsi dengan baik dan dapat menurunkan validitas instrumen secara keseluruhan. Kondisi ini sejalan dengan penelitian terbaru yang menekankan bahwa pengecoh yang tidak berfungsi dapat meningkatkan peluang peserta menebak jawaban (Ansari et al., 2022).

Kategori terakhir, *sangat buruk*, mencakup 8 butir soal (22,22%), menunjukkan pengecoh sama sekali tidak menarik respons peserta. Pengecoh yang gagal berfungsi secara konsisten menjadi indikator kuat perlunya revisi menyeluruh terhadap butir soal tersebut untuk meningkatkan kualitas pengukuran (Rameshbhai et al., 2023). Secara keseluruhan, hasil ini menegaskan bahwa sebagian besar butir memerlukan revisi agar instrumen memiliki kualitas pengecoh yang lebih kuat dan mampu menghasilkan asesmen yang valid serta reliabel.

Pembahasan

Hasil penelitian ini menunjukkan bahwa kualitas butir soal elemen komputer akuntansi fase F secara umum berada pada kategori baik, meskipun terdapat beberapa aspek yang memerlukan perbaikan signifikan. Analisis validitas memperlihatkan bahwa hampir separuh butir soal berada pada kategori *cukup*, diikuti dengan proporsi validitas tinggi yang relatif kuat. Temuan ini sejalan dengan penelitian Levacher et al. (2023) yang menekankan bahwa validitas konstruk tidak hanya dipengaruhi oleh kesesuaian isi, tetapi juga oleh bagaimana item mampu merefleksikan kemampuan aktual peserta didik. Banyaknya butir dalam kategori cukup menunjukkan perlunya revisi untuk memperkuat korelasi terhadap skor total, sehingga evaluasi pembelajaran dapat menghasilkan interpretasi yang lebih akurat.

Dari sisi reliabilitas, nilai reliabilitas tes sebesar 0,93 memberikan bukti bahwa instrumen memiliki konsistensi internal yang sangat baik. Hal ini menunjukkan bahwa sebagian besar butir mampu memberikan pengukuran yang stabil dan minim kesalahan. Menurut Jahrami et al. (2024), reliabilitas yang tinggi merupakan indikator bahwa varians skor lebih dipengaruhi oleh kemampuan

sebenarnya dibandingkan error measurement. Temuan ini mengindikasikan bahwa instrumen layak digunakan sebagai alat evaluasi pembelajaran akuntansi berbasis komputer.

Analisis tingkat kesukaran memperlihatkan bahwa mayoritas item berada pada kategori sedang, yang mencapai 47,22% dari keseluruhan butir. Kondisi ini ideal karena butir dengan tingkat kesukaran sedang umumnya mampu memberikan informasi yang paling optimal untuk mengukur kemampuan peserta dengan berbagai tingkat penguasaan (Zhang & Colvin, 2024). Namun, keberadaan sejumlah butir pada kategori sangat mudah dan sangat sukar mengindikasikan ketidakseimbangan tingkat tantangan pada sebagian aspek. Butir sangat mudah dapat mengurangi variasi respons peserta sehingga menurunkan daya beda, sedangkan butir sangat sukar dapat menyebabkan frustasi dan tidak mencerminkan kompetensi sebenarnya (Kim et al., 2024). Oleh karena itu, beberapa butir perlu disesuaikan agar kesukaran instrumen lebih proporsional.

Selanjutnya, analisis daya pembeda menunjukkan bahwa 41,67% butir berada pada kategori baik, dan 30,56% lainnya berkategori sangat baik. Ini merupakan indikator positif yang menunjukkan bahwa sebagian besar item mampu membedakan peserta berkemampuan tinggi dan rendah secara efektif. Daya pembeda yang tinggi sangat penting dalam penilaian berbasis tes, karena berperan dalam menilai ketepatan interpretasi hasil asesmen (Awalurahman & Budi, 2024). Meskipun demikian, ditemukannya butir dengan daya pembeda negatif menunjukkan bahwa terdapat item yang tidak hanya gagal berfungsi, tetapi juga memberikan sinyal evaluatif yang keliru, sehingga perlu diperbaiki atau dihapus (Kumar et al., 2021).

Analisis efektivitas pengecoh menunjukkan hasil yang lebih beragam. Sebanyak 27,78% butir berada pada kategori kurang baik, dan 22,22% lainnya masuk kategori sangat buruk. Temuan ini mengindikasikan bahwa sejumlah pengecoh tidak mampu menarik jawaban peserta didik secara proporsional. Pengecoh yang tidak berfungsi akan mengurangi efektivitas instrumen dalam menggambarkan kemampuan peserta dan dapat memengaruhi baik tingkat kesukaran maupun daya pembeda (Rezigalla et al., 2024). Rendahnya efektivitas pengecoh dalam beberapa butir dapat dipengaruhi oleh redaksi pengecoh yang terlalu jelas, tidak relevan, atau mudah dikenali sebagai jawaban salah (Lee et al., 2025). Oleh karena itu, revisi redaksional dan konseptual diperlukan untuk meningkatkan kualitas pengecoh.

Secara keseluruhan, integrasi hasil analisis menunjukkan bahwa instrumen soal sudah memenuhi sebagian besar aspek psikometris yang disyaratkan oleh teori tes klasik, terutama dari segi reliabilitas dan daya pembeda. Meskipun demikian, perbaikan diperlukan pada area validitas sedang-ke-bawah dan efektivitas pengecoh. Dengan memperbaiki item-item tersebut, kualitas instrumen akan meningkat, sehingga proses evaluasi pada pembelajaran komputer akuntansi dapat dilakukan secara lebih akurat dan representatif. Temuan ini juga menegaskan pentingnya penggunaan perangkat lunak analisis seperti Anates, yang terbukti mampu memberikan rekomendasi berbasis data empiris untuk meningkatkan mutu instrumen asesmen (Hermaya et al., 2024; Sabela et al., 2025).

Hasil penelitian ini diharapkan dapat menjadi dasar pengembangan instrumen asesmen yang lebih berkualitas di SMK, serta mendorong penelitian lanjutan untuk mengkaji karakteristik butir berdasarkan pendekatan modern seperti Item Response Theory (IRT) agar menghasilkan evaluasi yang lebih komprehensif di masa mendatang.

KESIMPULAN

Penelitian ini dilakukan untuk menganalisis kualitas butir soal elemen komputer akuntansi fase F di SMK dengan memanfaatkan perangkat lunak Anates berdasarkan lima indikator utama: validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Secara umum, hasil penelitian menunjukkan bahwa instrumen tes memiliki kualitas yang cukup baik, meskipun sejumlah aspek masih memerlukan perbaikan untuk memastikan pengukuran yang lebih akurat dan representatif terhadap kompetensi peserta didik. Analisis validitas menunjukkan bahwa sebagian besar butir berada pada kategori cukup, sementara proporsi butir berkategori tinggi dan sangat tinggi masih relatif terbatas. Hal ini mengindikasikan bahwa meskipun instrumen telah mencerminkan sebagian besar kompetensi yang diukur, beberapa butir item memerlukan revisi dari segi redaksi, relevansi indikator, maupun kesesuaian materi agar mampu menggambarkan konstruk secara lebih tepat.

Instrumen menunjukkan reliabilitas yang sangat tinggi (0,93), yang menegaskan bahwa tes memiliki konsistensi internal yang kuat dan mampu menghasilkan skor yang stabil antarpeserta. Dari sisi tingkat kesukaran, mayoritas butir tergolong sedang, sesuai karakteristik ideal dalam penyusunan tes. Namun, sejumlah butir sangat mudah dan sangat sukar masih perlu disesuaikan agar tidak menurunkan efektivitas pengukuran. Daya pembeda sebagian besar berada pada kategori baik hingga sangat baik, sehingga instrumen mampu membedakan siswa berkemampuan tinggi dan rendah dengan cukup efektif. Namun, keberadaan beberapa butir dengan daya pembeda negatif menunjukkan perlunya revisi signifikan. Pada aspek efektivitas pengecoh, sebagian besar pengecoh berada pada kategori kurang baik hingga sangat buruk, menandakan bahwa opsi jawaban belum bekerja optimal dalam mengalihkan peserta dari kunci jawaban Secara keseluruhan, instrumen tes elemen komputer akuntansi fase F memenuhi sebagian besar kriteria kualitas menurut teori tes klasik. Meski demikian, revisi komprehensif pada validitas moderat serta pengecoh tidak berfungsi diperlukan untuk meningkatkan mutu instrumen dan mendukung evaluasi pembelajaran kejuruan secara lebih akurat dan diagnostik.

DAFTAR PUSTAKA

A. Stephen Editor, L. (2024). Developments in the Reporting of Score Reliability within Counseling Assessment, Research, and Evaluation. *Measurement and Evaluation in Counseling and Development*, *57*(2), 89–92. https://doi.org/10.1080/07481756.2024.2333692

Agnola, E. B. Y., Aurelia, T. R., Hakim, L., & Pratiwi, V. (2025). Analisis Evaluasi Soal Komputer Akuntansi Menggunakan Software Anates oleh Siswa SMK Ketintang. *Jurnal Pendidikan Tambusai*, *9*(1), 2262–2270.

Ahmad, f., Kartika Dwi, A., Andini Kasih Agus, S., Luqman, H., & Vivi, P. (2024). Analisis Butir Soal HOTS Pilihan Ganda Pada Elemen Komputer Akuntansi Di SMKN 10 Surabaya Menggunakan Aplikasi Anates. *PESHUM : Jurnal Pendidikan, Sosial dan Humaniora*, *4*(1), 727–738. https://doi.org/10.56799/peshum.v4i1.6792

Alemu, A., Tesfa, H., Mulugeta, A., Fenta, E., & Belay, M. (2024). Quality of multiple choice question items: item analysis. *International Journal of Scientific Reports*, *10*, 195–199. https://doi.org/10.18203/issn.2454-2156.IntJSciRep20241316

Ansari, M., Sadaf, R., Akbar, A., Rehman, S., Chaudhry, Z., & Shakir, S. (2022). Assessment of distractor efficiency of MCQS in item analysis. *The Professional Medical Journal*, *29*, 730–734. https://doi.org/10.29309/TPMJ/2022.29.05.6955

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

Awalurahman, H., & Budi, I. (2024). Automatic distractor generation in multiple-choice questions: a systematic literature review. *PeerJ Computer Science*, *10*, e2441. https://doi.org/10.7717/peerj-cs.2441

Fiska, J., Hidayati, Y., Qomaria, N., & Puspita Hadi, W. (2021). Analisis Butir Soal Ulangan Harian IPA Menggunakan Software Anates Pada Pendekatan Teori Tes Klasik. *Natural Science Education Research*, *4*, 65–76. https://doi.org/10.21107/nser.v4i1.8133

Hermaya, I., Helendra, Vauzia, & Arsih, F. (2024). Item Quality Analysis of Concept Understanding and Problem Solving in Environmental Change Materials Using ANATES. *Jurnal Ilmiah Pendidikan Profesi Guru*, *7*, 371–381. https://doi.org/10.23887/jippg.v7i2.84083

Jahrami, H., Husain, W., Lin, C.-Y., Björling, G., Potenza, M. N., & Pakpour, A. (2024). Reliability generalization Meta-Analysis and psychometric review of the Gaming Disorder test (GDT): Evaluating internal consistency. *Addictive Behaviors Reports*, *20*, 100563. https://doi.org/https://doi.org/10.1016/j.abrep.2024.100563

Kim, Y. H., Kim, B. H., Kim, J., Jung, B., & Bae, S. (2024). Item difficulty index, discrimination index, and reliability of the 26 health professions licensing examinations in 2023, Korea: a psychometric study. *J Educ Eval Health Prof*, *21*, 40. https://doi.org/10.3352/jeehp.2024.21.40

Kreitchmann, R. S., Nájera, P., Sanz, S., & Sorrel, M. (2024). Enhancing Content Validity Assessment With Item Response Theory Modeling. *Psicothema*, *36*, 145–153. https://doi.org/10.7334/psicothema2023.208

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, *77*, S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Lee, Y., Kim, S., & Jo, Y. (2025). *Generating Plausible Distractors for Multiple-Choice Questions via Student Choice Prediction*.

Levacher, J., Koch, M., Stegt, S., Hissbach, J., Spinath, F., Escher, M., & Becker, N. (2023). The construct validity of the main student selection tests for medical studies in Germany. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1120129

Mardiyyaningsih, A., Erlina, E., Ulfah, M., & Wafiq, A. (2023). Validity and Reliability of the Two-tier Diagnostic Test to Identify Students' Alternative Conceptions of Intermolecular Forces. *Jurnal Penelitian Pendidikan IPA*, *9*, 4375–4381. https://doi.org/10.29303/jppipa.v9i6.2797

Marianti, S. R., Ana; Hasanah, Nur; and Nuryanti, Sofia. (2023). Comparing item-total correlation and item-theta correlation in test item selection: A simulation and empirical study. *Jurnal Penelitian dan Evaluasi Pendidikan*, *Vol. 27: Iss. 2, Article 1*. https://doi.org/DOI: 10.21831/pep.v27i2.61477

Mawardi, M., Fuady, A., & Sunismi, S. (2023). Analisis Butir Soal Pilihan Ganda Menggunakan Anates pada Penilaian Tengah Semester Kelas VII D SMP Negeri 1 Ngajum Kabupaten Malang. *Wahana*, *75*, 31–41. https://doi.org/10.36456/wahana.v75i1.6820

Mulyani, H., Tanuatmodjo, H., & Iskandar, R. (2020). Quality analysis of teacher-made tests in financial accounting subject at vocational high schools. *Jurnal Pendidikan Vokasi*, *10*. https://doi.org/10.21831/jpv.v10i1.29382

Nurjanah, S., Iqbal, M., Zafrullah, Z., Mahmud, M., Seran, D. a., Suardi, I., & Arriza, L. (2024). Psychometric quality of multiple-choice tests under Classical Test Theory (CTT): AnBuso, Iteman, and RStudio. *Jurnal Penelitian dan Evaluasi Pendidikan*, *28*. https://doi.org/10.21831/pep.v28i2.71542

Putri, R., Parwita, W., Handika, S., Sudipa, I. G. I., & Santika, P. (2024). Evaluation of Accounting Information System Using Usability Testing Method and System Usability Scale. *Sinkron*, *9*, 32–43. https://doi.org/10.33395/sinkron.v9i1.13129

Qurochman, A., Febriana, A., Santoso, H., & Haryana, R. (2024). Peningkatan Kompetensi Siswa SMK Melalui Pelatihan Aplikasi Komputer Accurate Accounting. *Jurnal Abmas Negeri (JAGRI)*, *5*, 494–503. https://doi.org/10.36590/jagri.v5i2.1311

Rameshbhai, C., Chauhan, B., Vaza, J., & Chauhan, P. (2023). Relations of the Number of Functioning Distractors With the Item Difficulty Index and the Item Discrimination Power in the Multiple Choice Questions. *Cureus*, *15*. https://doi.org/10.7759/cureus.42492

Rezigalla, A., Eleragi, A., Elhussein, A., Alfaifi, J., Alghamdi, M., Al-Ameer, A., Yahia, A., Mohamed, O., & Isa, A. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, *24*. https://doi.org/10.1186/s12909-024-05433-y

Sabela, O., Krisdayanty, D., Taqqiyah, A., Hakim, L., & Pratiwi, V. (2025). Analisis Butir Soal HOTS Elemen Dokumen Berbasis Digital (FASE E) Menggunakan Program Anates. *Education Achievement: Journal of Science and Research*, 251–262. https://doi.org/10.51178/jsr.v6i1.2328

Sheptian, R., Sarifah, I., & Riyadi, R. (2025). Classical Test Theory Analysis Using Anates: A Study of Mathematics Readiness Test for Elementary School Students. *SCIENCE : Jurnal Inovasi Pendidikan Matematika dan IPA*, *5*, 20–29. https://doi.org/10.51878/science.v5i1.3863

Sugiyono. (2022). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.

Sutrisno, P., Debora, D., Destriana, N., Putri, A., Marlinah, A., Wijaya, N., & Lekok, W. (2023). Pendampingan Pelatihan Software Akuntansi Accurate dalam Membantu Guru & Siswa-Siswi Smk untuk Meningkatkan Kompetensi dan Profesionalisme. *Jurnal Pemberdayaan Ekonomi*, *2*, 29–37. https://doi.org/10.35912/jpe.v2i1.716

Toma, R. B., Ortiz-Revilla, J., & Greca, I. (2024). Development and validation of a multiple-choice test for sustainability competence in primary school using the GreenComp framework. *International Journal of Educational Research Open*, *7*, 1–7. https://doi.org/10.1016/j.ijedro.2024.100388

Wilsa, A. W., Rusilowati, A., Susilaningsih, E., Jaja, J., & Nurpadillah, V. (2023). Validity, reliability, and item characteristics of cell material science literacy assessment instruments. *Jurnal Penelitian dan Evaluasi Pendidikan*, *27*(2), 177–188. https://doi.org/10.21831/pep.v27i2.61577

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), 126–135. https://doi.org/10.51878/learning.v4i1.2777

Zhang, S., & Colvin, K. (2024). Comparison of different reliability estimation methods for single-item assessment: a simulation study. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1482016