belantika Pendidikan
ISSN 2720-9601
Vol.8(2), November 2025, pp. 90-97

Analisis Kualitas Butir Soal Ujian Akhir Semester Pendidikan Agama Islam dan Budi Pekerti Kelas XII SMKN 3 Pontianak

Bayu Fitra Prisuna, Saumi Setyaningrum
Fakultas Tarbiyah dan Ilmu Keguruan, Institut Agama Islam Negeri Pontianak

INFO ARTIKEL

Riwayat Artikel:

Diterima: 06-11-2025
Disetujui: 27-12-2025

Kata kunci:

analisis butir soal;
validitas;
reliabilitas;
daya pembeda;
evaluasi pembelajaran

item analysis;
test quality;
difficulty level;
discriminating power;
distractor effectiveness;
learning evaluation

ABSTRAK

Abstract: This study aims to analyze the quality of the Even Semester Final Examination questions for Islamic Religious Education and Character Building subjects for class XII Accounting at SMKN 3 Pontianak based on validity, reliability, difficulty level, discriminating power, and distractor effectiveness. The study used a quantitative approach with a descriptive design. Data in the form of 40 multiple-choice questions and answer sheets from 49 students were analyzed using Anates version 4.0.9. The results showed that the quality of the instrument did not fully meet the criteria for a good test. Only 15% of the items were categorized as good, while most still needed revision. This finding demonstrates the importance of periodic item analysis to improve the quality of learning evaluation instruments.

Abstrak: Penelitian ini bertujuan menganalisis kualitas butir soal Ulangan Akhir Semester Genap mata pelajaran Pendidikan Agama Islam dan Budi Pekerti kelas XII Akuntansi SMKN 3 Pontianak berdasarkan validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Penelitian menggunakan pendekatan kuantitatif dengan desain deskriptif. Data berupa 40 soal pilihan ganda dan lembar jawaban 49 peserta didik dianalisis menggunakan Anates versi 4.0.9. Hasil penelitian menunjukkan bahwa kualitas instrumen belum sepenuhnya memenuhi kriteria tes yang baik. Hanya 15% butir soal berkategori baik, sedangkan sebagian besar masih memerlukan revisi. Temuan ini menunjukkan pentingnya analisis butir soal secara berkala untuk meningkatkan kualitas instrumen evaluasi pembelajaran.

Alamat Korespondensi:

Bayu Fitra Prisuna
Fakultas Tarbiyah dan Ilmu Keguruan
Institut Agama Islam Negeri Pontianak
Jl. Letnan Jenderal Soeprapto No. 19, Benua Melayu Darat, Kecamatan Pontianak Selatan, Kota Pontianak, Kalimantan Barat 78122
bayufitraprisuna@iainptk.ac.id

PENDAHULUAN

Evaluasi pembelajaran merupakan komponen fundamental dalam sistem pendidikan yang berfungsi untuk mengukur ketercapaian tujuan pembelajaran, menentukan tingkat penguasaan kompetensi peserta didik, serta menyediakan informasi yang diperlukan untuk perbaikan proses pembelajaran secara berkelanjutan (Ratnawulan & Rusdiana, 2015). Dalam praktik pendidikan, kualitas keputusan yang dihasilkan dari proses evaluasi sangat bergantung pada kualitas instrumen yang digunakan. Oleh karena itu, instrumen penilaian harus mampu menghasilkan data yang valid, reliabel, dan objektif agar interpretasi hasil belajar dapat dilakukan secara tepat dan akuntabel. Sejalan dengan *American Educational Research Association (AERA)*

dalam Zahner, (2025) standar internasional pengukuran pendidikan, kualitas instrumen menjadi dasar utama dalam menjamin validitas penggunaan hasil tes untuk berbagai kepentingan Pendidikan.

Tes hasil belajar, khususnya dalam bentuk pilihan ganda, masih menjadi instrumen yang paling banyak digunakan dalam evaluasi pembelajaran karena dinilai praktis, efisien, dan mampu mencakup cakupan materi yang luas. Namun, efektivitas suatu tes tidak hanya ditentukan oleh jumlah soal, melainkan oleh kualitas setiap butir yang menyusunnya. Analisis butir soal merupakan prosedur sistematis yang digunakan untuk mengevaluasi kualitas item berdasarkan indikator validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh (Pradani & Efendi, 2023). Penelitian terkini menunjukkan bahwa analisis butir soal yang komprehensif mampu memberikan informasi empiris mengenai performa setiap opsi jawaban, hubungan item dengan skor total, serta kualitas pengecoh yang digunakan sehingga dapat meningkatkan kualitas instrumen secara signifikan (Haladyna & Rodriguez, 2021).

Perkembangan penelitian pengukuran pendidikan dalam beberapa tahun terakhir juga menegaskan pentingnya analisis pengecoh *(distractor analysis)* dalam penyusunan soal pilihan ganda. Distraktor yang berfungsi dengan baik mampu meningkatkan daya pembeda dan kualitas psikometrik suatu item, sedangkan distraktor yang tidak efektif dapat menurunkan kemampuan soal dalam mengukur kompetensi peserta didik secara akurat. Penelitian Rezigalla et al., (2024) menemukan bahwa efisiensi pengecoh memiliki hubungan yang signifikan dengan indeks kesukaran dan daya pembeda butir soal, sehingga analisis distraktor perlu menjadi bagian integral dalam evaluasi kualitas tes. Selain itu, studi (Lions et al., 2021) menunjukkan bahwa karakteristik dan kualitas alternatif jawaban turut memengaruhi performa item dan validitas hasil pengukuran.

Meskipun penting, analisis butir soal masih belum menjadi praktik yang dilakukan secara rutin di banyak satuan pendidikan. Berdasarkan hasil wawancara awal dengan guru Pendidikan Agama Islam (PAI) di SMKN 3 Pontianak, diketahui bahwa soal Ulangan Akhir Semester (UAS) yang digunakan belum pernah dianalisis berdasarkan validitas, reliabilitas, tingkat kesukaran, daya pembeda, maupun efektivitas pengecoh. Akibatnya, kualitas instrumen yang digunakan untuk mengukur hasil belajar peserta didik belum diketahui secara objektif. Kondisi ini menunjukkan adanya kesenjangan antara praktik evaluasi yang dilaksanakan di sekolah dengan rekomendasi teoritis dan empiris yang menekankan pentingnya pengujian kualitas instrumen sebelum digunakan sebagai dasar pengambilan keputusan pendidikan. Padahal, instrumen yang tidak memenuhi standar kualitas berpotensi menghasilkan interpretasi yang bias terhadap kemampuan peserta didik dan mengurangi akurasi hasil evaluasi (Kreitzer & Sweet, 2022). Hal ini harus menjadi perhatian khusus bagi para guru dalam menyusun instrument tes yang baik, sehingga mampu mengukur apa yang hendak diukur. Menjawab tantangan tersebut maka sebagai pendidik profesional dituntut menjadi *problem solver* (Prisuna, 2021).

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan menganalisis kualitas butir soal Ulangan Akhir Semester Genap mata pelajaran Pendidikan Agama Islam dan Budi Pekerti kelas XII Akuntansi di SMKN 3 Pontianak Tahun Ajaran 2024/2025 berdasarkan aspek validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Penelitian ini memberikan kontribusi empiris terhadap pengembangan kajian evaluasi pembelajaran, khususnya pada mata pelajaran Pendidikan Agama Islam di jenjang sekolah menengah kejuruan. Selain itu, hasil penelitian diharapkan dapat menjadi dasar bagi guru dalam melakukan perbaikan instrumen penilaian secara berkelanjutan sehingga kualitas *asesment* yang digunakan dalam pembelajaran semakin sesuai dengan prinsip-prinsip pengukuran pendidikan modern.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan desain deskriptif. Pendekatan kuantitatif dipilih karena penelitian berfokus pada pengukuran karakteristik butir soal menggunakan data numerik yang dianalisis secara statistik, sedangkan desain deskriptif digunakan untuk menggambarkan kualitas instrumen evaluasi berdasarkan indikator yang telah ditetapkan tanpa melakukan manipulasi variabel atau pengujian hubungan antarvariabel.

Populasi penelitian mencakup seluruh dokumen evaluasi yang digunakan dalam pelaksanaan UAS Genap mata pelajaran Pendidikan Agama Islam dan Budi Pekerti kelas XII Akuntansi Tahun Ajaran 2024/2025. Sampel ditentukan menggunakan teknik *purposive sampling*, yaitu teknik pengambilan sampel berdasarkan pertimbangan tertentu yang sesuai dengan tujuan penelitian (Asrulla et al., 2023). Sampel penelitian terdiri atas 40 butir soal pilihan ganda beserta lembar jawaban 49 peserta didik yang berasal dari empat kelas, yaitu

XII AK 1, XII AK 2, XII AK 3, dan XII AK 4. Pemilihan sampel dilakukan karena dokumen tersebut merepresentasikan instrumen evaluasi yang digunakan secara resmi dalam penilaian hasil belajar peserta didik.

Instrumen penelitian berupa dokumen soal UAS, kunci jawaban, lembar jawaban peserta didik, serta perangkat lunak Anates versi 4.0.9 sebagai alat bantu analisis. Kualitas butir soal dianalisis berdasarkan lima indikator utama, yaitu validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh. Kelima indikator tersebut merupakan parameter yang umum digunakan dalam analisis butir soal berdasarkan pendekatan *Classical Test Theory (CTT)* untuk mengevaluasi kualitas instrumen tes Pendidikan (Fitriyah & Retnawati, 2023).

Data dikumpulkan menggunakan teknik dokumentasi melalui pengumpulan dan penelaahan dokumen yang berkaitan dengan pelaksanaan UAS, meliputi naskah soal, kunci jawaban, dan lembar jawaban peserta didik. Seluruh dokumen kemudian dikodekan dan dipersiapkan untuk proses analisis menggunakan perangkat lunak yang telah ditentukan. Selanjutnya analisis data dilakukan secara kuantitatif menggunakan perangkat lunak Anates versi 4.0.9. Analisis mencakup perhitungan koefisien validitas, reliabilitas tes, indeks tingkat kesukaran, daya pembeda, dan efektivitas pengecoh untuk setiap butir soal. Hasil analisis disajikan dalam bentuk statistik deskriptif berupa frekuensi, persentase, dan kategori kualitas butir soal berdasarkan kriteria yang berlaku. Selanjutnya, hasil tersebut diinterpretasikan untuk menentukan kelayakan setiap butir soal sebagai instrumen evaluasi pembelajaran. Pendekatan ini sejalan dengan rekomendasi penelitian pengukuran pendidikan yang menekankan pentingnya analisis empiris terhadap setiap item sebelum instrumen digunakan sebagai dasar pengambilan keputusan pendidikan (Rezigalla et al., 2024).

HASIL

Analisis kualitas butir soal dilakukan terhadap 40 soal pilihan ganda Ulangan Akhir Semester Genap mata pelajaran Pendidikan Agama Islam dan Budi Pekerti kelas XII Akuntansi SMKN 3 Pontianak Tahun Ajaran 2024/2025. Kualitas soal ditinjau berdasarkan validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan efektivitas pengecoh.

Interpretasi nilai validitas (korelasi) ditampilkan dalam tiga kategori, yakni "Sangat signifikan", "Signifikan" dan tanda strip (-) yang diartikan sebagai "Tidak signifikan". Hasil analisis menunjukkan bahwa sebagian besar butir soal belum memenuhi kriteria validitas yang baik. Berdasarkan 40 soal yang dianalisis, sebanyak 24 butir (60%) termasuk kategori tidak signifikan, sedangkan 5 butir (12,5%) berkategori signifikan dan 11 butir (27,5%) sangat signifikan (Gambar 1).

Korelasi Skor Butir dg Skor Total — Kembali Ke Menu Utama — Cetak

Jml Subyek= 49 Butir Soal = 40 Info tentang batas signifikansi

No Butir Baru	No Butir Asli	Korelasi	Signifikansi
1	1	0,429	Sangat Signifikan
2	2	0,372	Signifikan
3	3	-0,015	-
4	4	0,235	-
5	5	0,406	Sangat Signifikan
6	6	0,133	-
7	7	0,146	-
8	8	0,318	Signifikan
9	9	0,289	-
10	10	-0,036	-
11	11	-0,009	-
12	12	0,277	-
13	13	0,264	-
14	14	0,204	-
15	15	0,269	-
16	16	0,192	-
17	17	0,204	-
18	18	0,192	-
19	19	0,251	-
20	20	0,088	-
21	21	0,488	Sangat Signifikan
22	22	-0,066	-
23	23	0,469	Sangat Signifikan

Korelasi Skor Butir dg Skor Total — Kembali Ke Menu Utama — Cetak

Jml Subyek= 49 Butir Soal = 40 Info tentang batas signifikansi

No Butir Baru	No Butir Asli	Korelasi	Signifikansi
18	18	0,192	-
19	19	0,251	-
20	20	0,088	-
21	21	0,488	Sangat Signifikan
22	22	-0,066	-
23	23	0,469	Sangat Signifikan
24	24	0,132	-
25	25	0,279	-
26	26	0,467	Sangat Signifikan
27	27	0,189	-
28	28	0,467	Sangat Signifikan
29	29	0,312	Signifikan
30	30	0,523	Sangat Signifikan
31	31	0,011	-
32	32	0,078	-
33	33	0,346	Signifikan
34	34	0,346	Signifikan
35	35	0,580	Sangat Signifikan
36	36	0,523	Sangat Signifikan
37	37	0,162	-
38	38	0,237	-
39	39	0,424	Sangat Signifikan
40	40	0,480	Sangat Signifikan

Gambar 1. Output Validitas Butir Soal pada Anates Versi 4.0.9

Selanjutnya analisis reliabilitas butir soal pada Anates meliputi keseluruhan butir soal. Pengujian reliabilitas dimaksud untuk menguji sejauh mana konsistensi butir soal apabila diujikan di masa mendatang. Adapun nilai reliabilitas tes sebesar 0,39 menunjukkan bahwa instrumen memiliki tingkat konsistensi yang rendah. Hal ini dapat dilihat pada output yang dihasilkan dari software Anates sebagai berikut (Gambar 2).

Reliabilitas Tes

Reliabilitas Tes Kembali Ke Menu Utama Cetak

Rata2=28,51 Simpang Baku= 4,59 KorelasiXY= 0,24 Reliabilitas Tes = 0,39

No.Urut	No. Subyek	Kode/Nama Subyek	Skor Ganjil	Skor Genap	Skor Total
1	1	ACHMAD CHAIRUMMI	13	10	23
2	2	ANNISA APRILIA	16	11	27
3	3	CHERRY RICHELLA	12	11	23
4	4	DICKY AKMAL FAKHRI	17	10	27
5	5	DINA DALILAH	18	14	32
18	18	DELI FEBRIA DAMAYANT	18	15	33
19	19	FEBY	18	16	34
20	20	KEYLA APRILIA SABRINA	15	13	28
21	21	NADIA HUTAMI	14	11	25
22	22	NAJWA CAHYA NABILLA	9	9	18
23	23	NEYSHA PRAYENI	15	10	25

Reliabilitas Tes

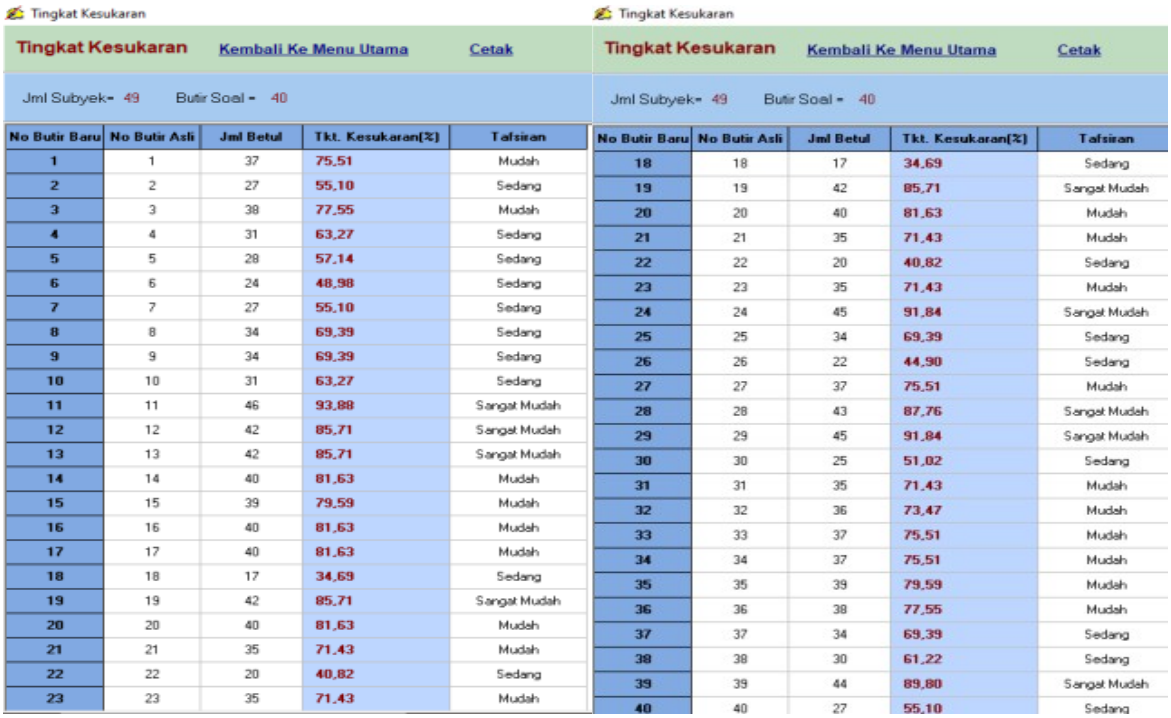
Reliabilitas Tes Kembali Ke Menu Utama Cetak

Rata2=28,51 Simpang Baku= 4,59 KorelasiXY= 0,24 Reliabilitas Tes = 0,39

No.Urut	No. Subyek	Kode/Nama Subyek	Skor Ganjil	Skor Genap	Skor Total
23	23	NEYSHA PRAYENI	15	10	25
24	24	RISKIA PURNAMA DEWI	15	10	25
25	25	SYALU SAHARANI	17	11	28
26	26	VARSHA ANCA SALSABILA	16	9	25
27	27	ADELIA	15	13	28
40	40	YURIKA PATRICIA	18	14	32
41	41	ALYA RIANTY PUTRI	12	18	30
42	42	ARDIANSYAH	8	9	17
43	43	AQSHAL	14	13	27
44	44	MELIA EKOL	9	16	25
45	45	NAHDA	12	10	22

Gambar 2. Output Reliabilitas Tes pada Anates Versi 4.0.9

Selanjutnya Interpretasi tingkat kesukaran dalam program Anates ditampilkan dalam tiga kategori, yakni “Sangat mudah”, “Mudah”, “Sedang”, “Sukar”, “Sangat sukar” (Gambar 3). Berdasarkan tinjauan dari tingkat kesukaran, sebagian besar soal berada pada kategori mudah hingga sedang. Sebanyak 24 butir (60%) tergolong mudah atau sangat mudah, sedangkan 16 butir (40%) berada pada kategori sedang. Tidak ditemukan soal yang tergolong sukar maupun sangat sukar seperti pada output Anates yang dapat dilihat sebagai berikut.



Tingkat Kesukaran

Tingkat Kesukaran Kembali Ke Menu Utama Cetak

Jml Subyek= 49 Butir Soal = 40

No Butir Baru	No Butir Asli	Jml Betul	Tkt. Kesukaran(%)	Tafsiran
1	1	37	75,51	Mudah
2	2	27	55,10	Sedang
3	3	38	77,55	Mudah
4	4	31	63,27	Sedang
5	5	28	57,14	Sedang
6	6	24	48,98	Sedang
7	7	27	55,10	Sedang
8	8	34	69,39	Sedang
9	9	34	69,39	Sedang
10	10	31	63,27	Sedang
11	11	46	93,88	Sangat Mudah
12	12	42	85,71	Sangat Mudah
13	13	42	85,71	Sangat Mudah
14	14	40	81,63	Mudah
15	15	39	79,59	Mudah
16	16	40	81,63	Mudah
17	17	40	81,63	Mudah
18	18	17	34,69	Sedang
19	19	42	85,71	Sangat Mudah
20	20	40	81,63	Mudah
21	21	35	71,43	Mudah
22	22	20	40,82	Sedang
23	23	35	71,43	Mudah

Tingkat Kesukaran

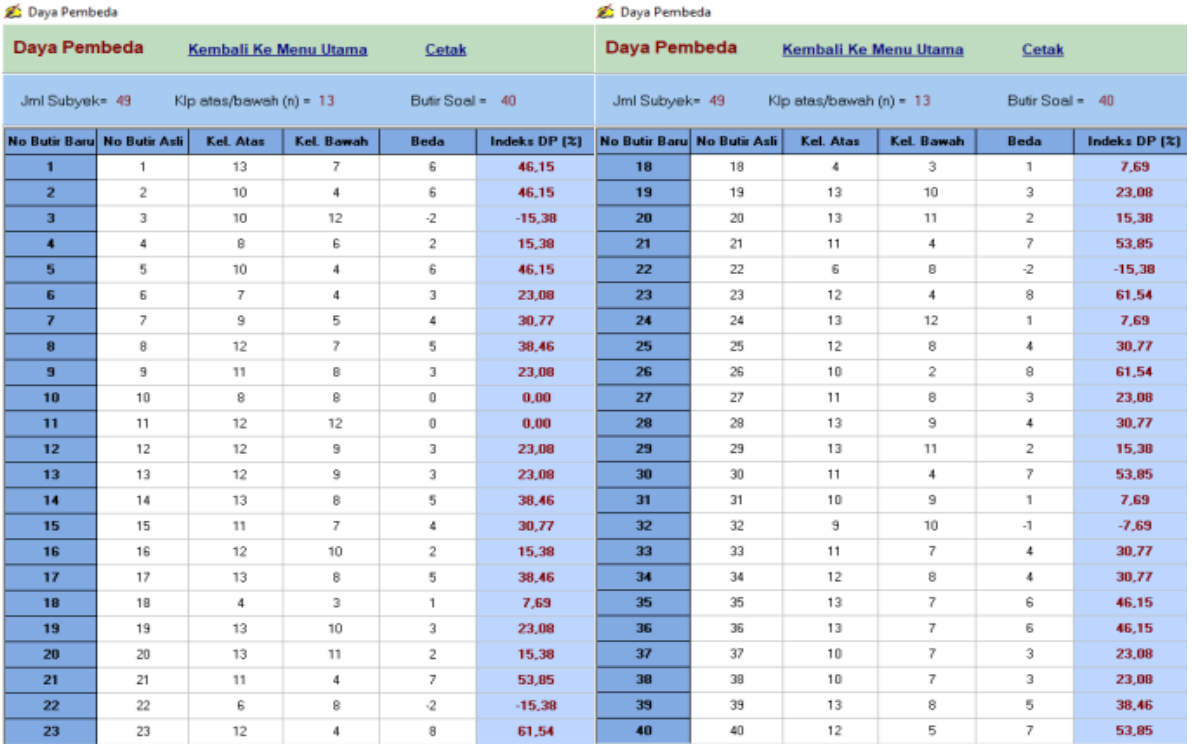
Tingkat Kesukaran Kembali Ke Menu Utama Cetak

Jml Subyek= 49 Butir Soal = 40

No Butir Baru	No Butir Asli	Jml Betul	Tkt. Kesukaran(%)	Tafsiran
18	18	17	34,69	Sedang
19	19	42	85,71	Sangat Mudah
20	20	40	81,63	Mudah
21	21	35	71,43	Mudah
22	22	20	40,82	Sedang
23	23	35	71,43	Mudah
24	24	45	91,84	Sangat Mudah
25	25	34	69,39	Sedang
26	26	22	44,90	Sedang
27	27	37	75,51	Mudah
28	28	43	87,76	Sangat Mudah
29	29	45	91,84	Sangat Mudah
30	30	25	51,02	Sedang
31	31	35	71,43	Mudah
32	32	36	73,47	Mudah
33	33	37	75,51	Mudah
34	34	37	75,51	Mudah
35	35	39	79,59	Mudah
36	36	38	77,55	Mudah
37	37	34	69,39	Sedang
38	38	30	61,22	Sedang
39	39	44	89,80	Sangat Mudah
40	40	27	55,10	Sedang

Gambar 3. Tingkat Kesukaran pada Anates Versi 4.0.9

Selanjutnya, berdasarkan hasil analisis daya pembeda butir soal Ulangan Akhir Semester Genap mata pelajaran Pendidikan Agama Islam dan Budi Pekerti Kelas XII SMKN 3 Pontianak Tahun Ajaran 2024/2025 melalui program Anates versi 4.0.9 terhadap 40 butir soal pilihan ganda menunjukkan bahwa sebanyak 9 soal atau 22,5% soal berkategori jelek, 18 soal atau 45% berkategori cukup, sebanyak 10 soal atau 25% berkategori baik, sebanyak 3 soal atau 7,5% berkategori negatif. Tidak terdapat soal yang berkategori baik sekali. Data tersebut dapat dilihat pada output Anates seperti Gambar 4.



Daya Pembeda

Daya Pembeda Kembali Ke Menu Utama Cetak

Jml Subyek= 49 Klp atas/bawah (n) = 13 Butir Soal = 40

No Butir Baru	No Butir Asli	Kel. Atas	Kel. Bawah	Beda	Indeks DP (%)
1	1	13	7	6	46,15
2	2	10	4	6	46,15
3	3	10	12	-2	-15,38
4	4	8	6	2	15,38
5	5	10	4	6	46,15
6	6	7	4	3	23,08
7	7	9	5	4	30,77
8	8	12	7	5	38,46
9	9	11	8	3	23,08
10	10	8	8	0	0,00
11	11	12	12	0	0,00
12	12	12	9	3	23,08
13	13	12	9	3	23,08
14	14	13	8	5	38,46
15	15	11	7	4	30,77
16	16	12	10	2	15,38
17	17	13	8	5	38,46
18	18	4	3	1	7,69
19	19	13	10	3	23,08
20	20	13	11	2	15,38
21	21	11	4	7	53,85
22	22	6	8	-2	-15,38
23	23	12	4	8	61,54

Daya Pembeda

Daya Pembeda Kembali Ke Menu Utama Cetak

Jml Subyek= 49 Klp atas/bawah (n) = 13 Butir Soal = 40

No Butir Baru	No Butir Asli	Kel. Atas	Kel. Bawah	Beda	Indeks DP (%)
18	18	4	3	1	7,69
19	19	13	10	3	23,08
20	20	13	11	2	15,38
21	21	11	4	7	53,85
22	22	6	8	-2	-15,38
23	23	12	4	8	61,54
24	24	13	12	1	7,69
25	25	12	8	4	30,77
26	26	10	2	8	61,54
27	27	11	8	3	23,08
28	28	13	9	4	30,77
29	29	13	11	2	15,38
30	30	11	4	7	53,85
31	31	10	9	1	7,69
32	32	9	10	-1	-7,69
33	33	11	7	4	30,77
34	34	12	8	4	30,77
35	35	13	7	6	46,15
36	36	13	7	6	46,15
37	37	10	7	3	23,08
38	38	10	7	3	23,08
39	39	13	8	5	38,46
40	40	12	5	7	53,85

Gambar 4. Output Daya Pembeda pada Anates Versi 4.0.9.

Selanjutnya pada aspek efektivitas pengecoh yang dilakukan dengan program Anates versi 4.0.9 diperoleh hasil bahwa dari 40 butir soal yang dianalisis, terdapat 9 soal atau sebesar 22,5% berkategori sangat baik, 10 soal atau sebesar 25% sebesar berkategori baik, 14 soal atau sebesar 35% berkategori cukup, 5 soal atau sebesar 12,5% berkategori kurang baik, dan sebanyak 2 soal atau sebesar 5% berkategori tidak baik. Data tersebut dapat dilihat pada output Anates pada Gambar 5.

Kualitas Pengecoh

Jml Subyek= 49 Butir Soal = 40 ** : Kunci Jawaban +: Baik -- : Buruk ++ : Sangat Baik -: Kurang --- : Sangat Buruk

Kualitas Pengecoh Kembali Ke Menu Utama Cetak

No Butir Baru	No Butir Asli	a	b	c	d	e	*
1	1	6---	4+	0--	37**	0--	0
2	2	4+	13*	27**	3+	0--	0
3	3	9---	0--	30**	2+	0--	0
4	4	4++	2-	0--	31**	12---	0
5	5	0-	28**	3-	1--	9-	0
6	6	0--	11--	5++	9+	24**	0
7	7	27**	6++	3-	10---	3+	0
8	8	2+	2+	5-	5+	34**	0
9	9	34**	2+	7--	3++	2+	0
10	10	2-	6+	3-	31*	6+	0
11	11	2---	48**	0--	0--	0--	0
12	12	2++	3-	42**	2++	0--	0
13	13	3-	2++	2++	42**	0--	0
14	14	3+	2++	40**	3+	1-	0
15	15	4-	2++	2++	39**	2++	0
16	16	5---	1-	0--	3+	40**	0
17	17	48**	2---	1-	1-	0--	0
18	18	9++	14-	5+	3-	17**	0
19	19	42**	4---	2++	1+	0--	0
20	20	4--	40**	3+	1-	1-	0
21	21	2+	5+	4++	3++	35**	0
22	22	7++	0--	1--	21---	20**	0
23	23	0--	5+	1-	38**	8---	0

Kualitas Pengecoh Kembali Ke Menu Utama Cetak

No Butir Baru	No Butir Asli	a	b	c	d	e	*
18	18	9++	14-	5+	3-	17**	0
19	19	42**	4---	2++	1+	0--	0
20	20	4--	40**	3+	1-	1-	0
21	21	2+	5+	4++	3++	35**	0
22	22	7++	0--	1--	21---	20**	0
23	23	0--	5+	1-	38**	8---	0
24	24	1++	45**	0--	0--	3---	0
25	25	4++	8---	34**	3++	0--	0
26	26	16---	7++	1--	3-	22**	0
27	27	2+	1-	37**	0--	3---	0
28	28	0--	3--	43**	3--	0--	0
29	29	0--	3---	1++	45**	0--	0
30	30	5++	5++	1--	13---	25**	0
31	31	2+	3++	3++	35**	6-	0
32	32	3++	3++	4++	36**	3++	0
33	33	37**	1-	7---	3++	1-	0
34	34	7---	37**	1-	4+	0--	0
35	35	4-	39**	1-	5--	0--	0
36	36	38**	1-	2+	5--	3++	0
37	37	7---	2+	6-	34**	0--	0
38	38	30**	3+	2-	2-	12---	0
39	39	0--	3---	1++	1++	44**	0
40	40	2-	27**	3+	9-	4+	0

Gambar 5. Efektivitas Pengecoh pada Anates Versi 4.0.9.

Berdasarkan sintesis seluruh indikator kualitas soal, diperoleh 6 butir soal (15%) berkategori baik, 21 butir soal (52,5%) berkategori cukup baik, dan 13 butir soal (32,5%) berkategori tidak baik.

PEMBAHASAN

Hasil penelitian menunjukkan bahwa kualitas instrumen Ulangan Akhir Semester Genap mata pelajaran Pendidikan Agama Islam dan Budi Pekerti di SMKN 3 Pontianak masih memerlukan perbaikan pada beberapa aspek utama, terutama validitas, reliabilitas, tingkat kesukaran, dan daya pembeda. Temuan ini mengindikasikan bahwa sebagian besar butir soal belum mampu berfungsi secara optimal sebagai alat ukur capaian belajar peserta didik. Dalam perspektif pengukuran pendidikan, kualitas instrumen merupakan prasyarat penting untuk menghasilkan informasi yang akurat dan dapat digunakan sebagai dasar pengambilan keputusan pendidikan. Menurut *Standards for Educational and Psychological Testing*, validitas merupakan dasar utama dalam mengevaluasi kualitas instrumen karena berkaitan dengan ketepatan interpretasi dan penggunaan skor tes Zahner, (2025). Oleh karena itu, rendahnya kualitas beberapa indikator psikometrik pada instrumen yang diteliti berpotensi memengaruhi akurasi informasi yang dihasilkan mengenai penguasaan kompetensi peserta didik.

Temuan penelitian menunjukkan bahwa 60% butir soal tidak memenuhi kriteria validitas. Persentase tersebut mengindikasikan bahwa sebagian besar soal belum mampu mengukur kompetensi yang hendak diukur secara tepat. Secara teoritis, validitas mencerminkan sejauh mana bukti empiris dan teori mendukung interpretasi skor tes sesuai tujuan penggunaannya. Instrumen yang tidak valid berpotensi menghasilkan kesalahan interpretasi terhadap kemampuan peserta didik sehingga keputusan yang diambil berdasarkan hasil tes menjadi kurang akurat. Temuan ini menunjukkan bahwa proses penyusunan soal kemungkinan belum sepenuhnya memperhatikan kesesuaian antara indikator pembelajaran, tujuan evaluasi, dan konstruksi item. Kondisi tersebut juga dapat disebabkan oleh redaksi soal yang ambigu, ketidaksesuaian materi yang diujikan dengan indikator kompetensi, atau kualitas alternatif jawaban yang kurang memadai. Penelitian Mannan & Kusaeri, (2025) menegaskan bahwa validitas instrumen merupakan aspek fundamental dalam evaluasi pendidikan karena menentukan kualitas inferensi yang dapat dibuat dari hasil pengukuran. Oleh karena itu, analisis validitas perlu dilakukan secara sistematis sebelum instrumen digunakan dalam evaluasi pembelajaran.

Rendahnya validitas soal tampak konsisten dengan hasil reliabilitas tes yang hanya mencapai koefisien 0,39. Nilai tersebut menunjukkan bahwa konsistensi hasil pengukuran masih berada pada kategori rendah. Dalam teori pengukuran klasik *(Classical Test Theory),* reliabilitas mengacu pada tingkat kestabilan dan konsistensi skor yang dihasilkan oleh suatu instrumen. Instrumen dengan reliabilitas rendah cenderung menghasilkan skor yang dipengaruhi oleh kesalahan pengukuran *(measurement error)* sehingga kurang mampu merepresentasikan kemampuan peserta didik secara konsisten. Hubungan antara validitas dan reliabilitas telah lama menjadi perhatian dalam penelitian psikometri. Instrumen yang memiliki banyak butir tidak valid umumnya akan menghasilkan reliabilitas yang rendah karena skor total tidak mencerminkan konstruk yang diukur secara konsisten. Temuan penelitian ini berbeda dengan hasil penelitian Saptaputra et al., (2023) yang menunjukkan bahwa instrumen evaluasi PAI yang dianalisis memiliki reliabilitas tinggi sehingga layak digunakan sebagai alat ukur hasil belajar. Perbedaan tersebut mengindikasikan bahwa kualitas instrumen sangat dipengaruhi oleh proses pengembangan dan evaluasi butir soal yang dilakukan sebelum instrumen digunakan.

Berdasarkan aspek tingkat kesukaran, penelitian ini menemukan bahwa sebagian besar soal berada pada kategori mudah dan sedang, sementara tidak ditemukan soal yang tergolong sukar. Temuan ini menunjukkan bahwa distribusi tingkat kesukaran belum proporsional. Menurut Ratnawulan & Rusdiana, (2015), instrumen yang baik seharusnya memiliki komposisi tingkat kesukaran yang seimbang agar mampu mengukur kemampuan peserta didik pada berbagai tingkatan. Dominasi soal mudah dapat menyebabkan skor peserta didik cenderung terkonsentrasi pada nilai tinggi sehingga variasi kemampuan sulit dibedakan secara akurat. Dalam konteks evaluasi pembelajaran, soal dengan tingkat kesukaran sedang umumnya memberikan informasi yang lebih optimal karena mampu mengidentifikasi perbedaan kemampuan peserta didik secara lebih jelas. Hasil penelitian ini sejalan dengan temuan Shafara et al., (2024) yang menunjukkan bahwa dominasi soal mudah dapat menurunkan efektivitas tes dalam membedakan peserta didik berkemampuan tinggi dan rendah.

Selain itu, tidak adanya soal kategori sukar mengindikasikan bahwa instrumen kurang sensitif dalam mengidentifikasi peserta didik yang memiliki penguasaan materi pada tingkat yang lebih kompleks.

Kondisi tersebut terlihat konsisten dengan hasil analisis daya pembeda. Penelitian ini menemukan bahwa hanya seperempat butir soal yang memiliki daya pembeda baik, sedangkan sebagian besar berada pada kategori cukup dan jelek. Bahkan, terdapat tiga butir soal yang memiliki daya pembeda negatif. Menurut Nurhalimah et al., (2022), daya pembeda negatif menunjukkan bahwa peserta didik dengan kemampuan rendah justru lebih banyak menjawab benar dibandingkan peserta didik dengan kemampuan tinggi. Fenomena ini mengindikasikan adanya masalah serius dalam kualitas item, seperti kesalahan kunci jawaban, redaksi soal yang membingungkan, atau ketidaksesuaian antara soal dan indikator kompetensi. Dalam perspektif pengukuran pendidikan, daya pembeda merupakan indikator penting karena menunjukkan kemampuan suatu item dalam mengidentifikasi variasi kemampuan peserta didik. Haladyna & Rodriguez, (2021) menegaskan bahwa item yang memiliki daya pembeda tinggi berkontribusi terhadap peningkatan kualitas tes secara keseluruhan karena mampu memberikan informasi yang lebih akurat mengenai perbedaan kemampuan individu. Oleh sebab itu, temuan rendahnya daya pembeda pada sebagian besar butir soal menunjukkan perlunya revisi substansial terhadap kualitas item yang digunakan.

Berbeda dengan aspek validitas, reliabilitas, dan daya pembeda, efektivitas pengecoh menunjukkan hasil yang relatif baik. Sebanyak 82,5% butir soal memiliki pengecoh yang berfungsi pada kategori cukup baik hingga sangat baik. Temuan ini menunjukkan bahwa sebagian besar alternatif jawaban mampu menarik peserta didik yang belum menguasai materi sehingga berfungsi sebagaimana mestinya. Dalam soal pilihan ganda, distraktor yang efektif berperan penting dalam meningkatkan kualitas psikometrik suatu item. Distraktor yang tidak dipilih oleh peserta didik menunjukkan bahwa alternatif tersebut tidak berfungsi dan perlu direvisi. Penelitian Rezigalla et al., (2024) menunjukkan bahwa efektivitas distraktor memiliki hubungan yang signifikan dengan tingkat kesukaran dan daya pembeda suatu item. Semakin baik fungsi distraktor, semakin besar peluang suatu butir soal memiliki kemampuan membedakan peserta didik secara akurat. Temuan penelitian ini menunjukkan bahwa meskipun kualitas instrumen secara umum masih perlu ditingkatkan, aspek penyusunan alternatif jawaban telah menunjukkan kualitas yang relatif memadai dan dapat menjadi modal awal dalam pengembangan instrumen yang lebih baik.

Secara keseluruhan, hanya 15% butir soal yang memenuhi kategori baik dan layak digunakan tanpa revisi. Sementara itu, lebih dari setengah jumlah soal masih memerlukan perbaikan sebelum dapat dimasukkan ke dalam bank soal sekolah. Temuan ini mengindikasikan bahwa proses penyusunan instrumen evaluasi belum sepenuhnya didasarkan pada analisis empiris kualitas butir soal. Padahal, evaluasi pembelajaran modern menempatkan analisis butir soal sebagai bagian integral dari siklus penjaminan mutu asesmen. Savika & Zuhriyah, (2024) menegaskan bahwa analisis butir soal memungkinkan guru memperoleh umpan balik objektif mengenai kualitas instrumen yang digunakan sehingga perbaikan dapat dilakukan secara berkelanjutan. Temuan penelitian ini juga memperkuat rekomendasi AERA, APA, dan NCME bahwa penggunaan instrumen pendidikan seharusnya didasarkan pada bukti validitas dan reliabilitas yang memadai agar hasil evaluasi dapat digunakan secara bertanggung jawab.

Secara teoretis, penelitian ini memperkuat pentingnya analisis butir soal sebagai mekanisme evaluasi kualitas instrumen dalam kerangka *Classical Test Theory*. Secara praktis, hasil penelitian memberikan informasi empiris bagi guru Pendidikan Agama Islam untuk memperbaiki kualitas instrumen evaluasi melalui revisi item, penyusunan bank soal, dan pelaksanaan analisis butir soal secara berkala. Dengan demikian, instrumen evaluasi yang digunakan tidak hanya memenuhi standar teknis pengukuran, tetapi juga mampu mendukung pengambilan keputusan pendidikan yang lebih akurat, objektif, dan berorientasi pada peningkatan kualitas pembelajaran

SIMPULAN

Berdasarkan hasil penelitian dan pembahasan maka dapat disimpulkan bahwa bahwa kualitas butir soal Ulangan Akhir Semester Genap mata pelajaran Pendidikan Agama Islam dan Budi Pekerti kelas XII Akuntansi SMKN 3 Pontianak Tahun Ajaran 2024/2025 belum sepenuhnya memenuhi kriteria instrumen evaluasi yang baik. Dari 40 butir soal yang dianalisis, sebagian besar belum valid, memiliki reliabilitas rendah (0,39), didominasi tingkat kesukaran mudah dan sedang tanpa adanya soal sukar, serta memiliki daya pembeda yang sebagian besar berada pada kategori cukup dan jelek, meskipun efektivitas pengecoh menunjukkan hasil yang

relatif baik. Temuan ini menegaskan pentingnya analisis butir soal secara sistematis sebelum instrumen digunakan dalam evaluasi pembelajaran. Oleh karena itu, guru perlu melakukan analisis kualitas soal secara berkala untuk memperbaiki atau mengganti butir yang kurang berkualitas, sekolah perlu memperkuat penjaminan mutu asesmen melalui pengembangan bank soal dan pelatihan penyusunan instrumen berbasis analisis empiris, sedangkan penelitian selanjutnya disarankan memperluas cakupan sampel dan menggunakan pendekatan pengukuran yang lebih mutakhir, seperti *Item Response Theory (IRT)*, untuk memperoleh informasi psikometrik yang lebih komprehensif.

DAFTAR RUJUKAN

Asrulla, Risnita, Jailani, M. S., & Jeka, F. (2023). Populasi dan Sampling (Kuantitatif), Serta Pemilihan Informan Kunci (Kualitatif) dalam Pendekatan Praktis. *Jurnal Pendidikan Tambusai*, *7*(3), 26320–26332. https://doi.org/https://doi.org/10.31004/jptam.v7i3.10836

Fitriyah, I. M., & Retnawati, H. (2023). Analysis of the distractor of the multiple-choice test using classical test theory (CTT) and item response theory (IRT). *Materials of International Practical Internet Conference "Challenges of Science," Vi*. https://doi.org/https://doi.org/10.31643/2023.23

Haladyna, T. M., & Rodriguez, M. C. (2021). Using full-information item analysis to improve item quality. *Educational Assessment*, *26*(3), 198–211. https://doi.org/https://doi.org/10.1080/10627197.2021.1946390

Kreitzer, R. J., & Sweet, J. (2022). Evaluating Student Evaluations of Teaching : a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform. *Journal of Academic Ethics*, 73–84. https://doi.org/10.1007/s10805-021-09400-w

Lions, S., Monsalve, C., Dartnell, P., Godoy, M. I., & Córdova, N. (2021). The Position of Distractors in Multiple-Choice Test Items : The Strongest Precede the Weakest. *Front. Educ.*, *6*(October), 1–6. https://doi.org/10.3389/feduc.2021.731763

Mannan, A., & Kusaeri, S. (2025). Practices and Challenges of the Validity of Exploratory Factor Analysis (EFA) -Based Assessment Instruments : A Systematic Literature Review 2020-2025. *EDUKASIA: Jurnal Pendidikan Dan Pembelajaran*, *6*(1), 611–626. https://doi.org/10.62775/edukasia.v6i1.1463

Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). Hubungan Antara Validitas Item dengan Daya Pembeda dan Tingkat Kesukaran Soal Pilihan Ganda PAS. *Jurnal Natural Science Educational Research*, *4*(3), 249–257. https://doi.org/DOI: https://doi.org/10.21107/nser.v4i3.8682

Pradani, R. A., & Efendi, A. (2023). Analisis Butir Soal Ujian Sekolah Menggunakan Program Iteman (Analysis of School Exam Questions Using the Iteman Program). *Indonesian Language Education and Literature*, *8*(2), 276–289. https://doi.org/10.24235/ileal.v8i2.11002

Prisuna, B. F. (2021). *Pengaruh Penggunaan Aplikasi Google Meet terhadap Hasil Belajar The Effect of Using Google Meet Applications on Learning Outcomes*. *14*(2), 137–147. https://doi.org/10.21831/jpipfip.v14i1.39160

Ratnawulan, E., & Rusdiana, A. (2015). *Evaluasi pembelajaran*. CV Pustaka Setia.

Rezigalla, A. A., Mohammed, A., Seid, E., Eleragi, A., Elhussein, A. B., Alfaifi, J., Alghamdi, M. A., Ameer, A. Y. Al, Ibrahim, A., Yahia, O., Mohammed, O. A., Ishag, M., & Adam, E. (2024). Item analysis : the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*, *24*, 1–7. https://doi.org/https://doi.org/10.1186/s12909-024-05433-y

Saptaputra, I., Marwiyah, S., & Hasanuddin, M. I. (2023). Analysis of the Items Final Semester Test of Islamic Religious Education. *IQRO: Journal of Islamic Education*, *6*(1), 1–12. https://doi.org/DOI: https://doi.org/10.24256/iqro.v6i1.2391

Savika, H. I., & Zuhriyah, I. A. (2024). Peran analisis butir soal terhadap kualitas soal, kompetensi guru, dan prestasi belajar peserta didik di sekolah dasar. *Jurnal Pendidikan Anak Dan Pendidikan Umum*, *2*(2), 43–51. https://doi.org/10.59966/pandu.v2i2.856

Shafara, N. I., Ihsanudin, & Rafianti, I. (2024). Analisis Kemampuan Numerasi Matematis Peserta Didik Dalam Menyelesaikan Soal Asesmen Kompetensi Minimum. *Jurnal Educatio*, *10*(2), 614–622. https://doi.org/DOI : https://doi.org/10.31949/educatio.v10i2.8840

Zahner, D., Steedle, J. T., Soland, J., Welch, C., Qin, Q., Thompson, K., & Phelps, R. (2025). Results from NCME Survey on Revisions to the Standards for Educational and Psychological Testing. In *EdWorkingPaper: 25 -1124* (Issue 25). https://doi.org/https://doi.org/10.26300/7zzf-9193