

Item Analysis of Arabic Language Questions of Students in Grade XI MA Islamiyah Sidoarjo Based on Anates Software Version 4.0.9

[Analisis Butir Soal Bahasa Arab Siswa Kelas XI MA Islamiyah Sidoarjo Berbasis Software Anates Versi 4.0.9]

Yumna Hanifa Abdillah¹⁾, Khizanatul Hikmah^{*2)}

¹⁾Program Studi Pendidikan Bahasa Arab, Universitas Muhammadiyah Sidoarjo, Indonesia

²⁾ Program Studi Pendidikan Bahasa Arab, Universitas Muhammadiyah Sidoarjo, Indonesia

*Email Penulis Korespondensi: khizanatul.hikmah@umsida.ac.id

Abstract. This study aims to determine the quality of Arabic language questions for grade XI students of MA Islamiyah Sidoarjo using Anates software version 4.0.9. This research uses quantitative descriptive method with research subjects as many as 33 students of class XI IPA. The data collection technique used by researchers is documentation. The data in the form of 10 multiple-choice questions for the odd semester mid-summative were analyzed based on five criteria, namely validity, reliability, difficulty level discriminating power, and quality of distractors. The results showed that 60% of the questions were valid and 40% were invalid. The reliability value reached 0.90 indicating high consistency. In terms of difficulty level, 50% of the questions were in the easy category, 30% were very easy, and 20% were moderate, with no questions in the difficult category. For discriminating power, 40% of the questions had excellent differentiating power, 20% were good, 30% were sufficient, and 10% were not good. Meanwhile, 70% of distractors are considered to be functioning well, while 30% needed to be repaired. This study provides recommendations for teachers to improve the quality of questions by correcting items that do not meet the criteria, so that learning evaluations can be more effective in measuring student competence.

Keywords - item analysis; arabic language; anates software

Abstrak. Penelitian ini bertujuan untuk mengetahui kualitas butir soal bahasa Arab siswa kelas XI MA Islamiyah Sidoarjo dengan menggunakan software Anates versi 4.0.9. Penelitian ini menggunakan metode deskriptif kuantitatif dengan subjek penelitian sebanyak 33 siswa kelas XI IPA. Teknik pengumpulan data yang digunakan peneliti adalah dokumentasi. Data berupa 10 butir soal pilihan ganda Sumatif Tengah Semester ganjil yang dianalisis berdasarkan lima kriteria yaitu validitas, reliabilitas, tingkat kesukaran, daya pembeda, dan kualitas pengecoh (distractor). Hasil penelitian menunjukkan bahwa 60% soal tergolong valid dan 40% soal tidak valid. Nilai reliabilitas mencapai 0,90 yang menunjukkan konsistensi yang tinggi. Dari segi tingkat kesukaran, 50% soal termasuk dalam kategori mudah, 30% sangat mudah, dan 20% sedang, tanpa adanya soal yang berkategori sukar. Untuk daya pembeda, 40% soal memiliki daya pembeda yang sangat baik, 20% baik, 30% cukup, dan 10% tidak baik. Sementara itu, terdapat 70% distraktor dianggap berfungsi dengan baik, sedangkan 30% perlu diperbaiki. Penelitian ini memberikan rekomendasi kepada guru untuk meningkatkan kualitas soal dengan memperbaiki butir soal yang tidak memenuhi kriteria, sehingga evaluasi pembelajaran dapat lebih efektif dalam mengukur kompetensi siswa.

Kata Kunci - analisis butir soal; bahasa arab; software anates

I. PENDAHULUAN

Kualitas pendidikan tidak terlepas dari kualitas evaluasi yang diterapkan. Dalam sistem pendidikan manapun, evaluasi menjadi jendela untuk mengukur kualitas proses pendidikan. Sehingga, evaluasi menjadi bagian penting yang tidak dapat terpisahkan dari proses pendidikan, juga merupakan pilar dasar dari sistem pendidikan [1]. Bagaimana tidak, tujuan pendidikan yang ingin dicapai akan bergantung pada hasil evaluasi yang dilakukan. Metode evaluasi akan merujuk pada tujuan dari pendidikan tersebut. Oleh sebab itu, evaluasi menjadi upaya terorganisir untuk menilai sistem pembelajaran tertentu. Sejauh mana siswa telah mencapai tujuan pendidikan. Hal ini mencakup pengumpulan informasi, analisis informasi tersebut, dan serangkaian standar yang sesuai dengan target kompetensi [2]. Dengan demikian, evaluasi mampu berkontribusi pada peningkatan kualitas pembelajaran.

Tingkat keberhasilan pembelajaran akan diketahui melalui proses evaluasi hasil belajar siswa. Proses ini diharapkan dapat mengukur pertumbuhan dan perkembangan siswa [1]. Proses pembelajaran di sekolah dilengkapi dengan visi misi dan tujuan yang harus diraih. Tujuan tersebut dilakukan berupa evaluasi yang akan mengukur capaian pembelajaran siswa terhadap materi yang telah disampaikan. Oleh karena itu, evaluasi memiliki peran penting untuk

menilai seberapa jauh capaian pembelajaran dan tujuan pendidikan yang telah dicapai. [2]. Proses ini akan mengukur tingkat keberhasilan dan efisiensi pembelajaran yang telah dilaksanakan.

Selain itu, evaluasi juga sebagai alat penting untuk mengukur seberapa mudah peserta didik atau siswa menyerap pelajaran [3]. Dalam proses ini pendidik memiliki peranan untuk meninjau perkembangan peserta didik selama proses pembelajaran berlangsung. Pendidik atau guru harus dapat merencanakan, membentuk, dan membuat alat penilaian [4]. Ini berarti evaluasi hasil belajar siswa dilakukan setiap kali pendidik memberikan pelajaran, tidak hanya sekali dalam satu semester. Sementara itu, Magdalena juga menerangkan bahwa kegiatan evaluasi pembelajaran adalah untuk mendapatkan pemahaman yang menyeluruh tentang bagaimana atau apa yang berhasil dan gagal dalam proses pembelajaran. Juga untuk menemukan langkah-langkah perbaikan yang diperlukan untuk mencapai hasil pembelajaran yang lebih baik [5]. Dengan demikian, evaluasi sangatlah penting untuk mengetahui apakah proses pembelajaran berhasil dan apakah tujuan yang ditetapkan tercapai [6].

Evaluasi hasil belajar siswa oleh pendidik dapat menggunakan salah satu alat tes. Menurut Mulyadi, tes merupakan instrumen atau prosedur berupa pertanyaan yang harus dijawab, tugas yang harus dilakukan, baik secara lisan maupun tulisan [7]. Berdasarkan fungsinya, tes dalam pembelajaran bahasa Arab terbagi menjadi empat: *pertama* adalah tes penempatan, *kedua* tes formatif, *ketiga* tes diagnostik, dan *yang keempat* tes sumatif [1]. Tes kompetensi kebahasaan adalah alat bantu yang dipergunakan untuk mengevaluasi pemahaman dan penguasaan kebahasaan seseorang, terutama tes bahasa Arab [8]. Tes yang baik akan berdampak pada kualitas pembelajaran siswa. Maka dari itu, soal-soal harus dianalisis untuk mengetahui kualitasnya, apakah sudah memenuhi kriteria [9]. Seringkali, guru tidak menganalisis item soal yang digunakan untuk tes, sehingga guru tidak mengetahui kualitas soal yang dibuat.

Analisis butir soal merupakan hal yang penting untuk mengetahui kualitas soal yang dibuat. Proses ini mengevaluasi kualitas setiap butir soal dengan tujuan untuk menentukan soal yang baik, kurang baik, atau jelek. Setelah mengetahui kualitas setiap soal, berikutnya akan diberikan instruksi untuk memperbaiki soal-soal yang belum terpenuhi kriterianya, sehingga soal menjadi berkualitas [10]. Analisis butir soal memungkinkan untuk mengidentifikasi pertanyaan mana yang paling banyak dijawab oleh siswa dengan benar dan sebaliknya. Melalui sekolah yang akan diteliti ini, memungkinkan pengajar yang memiliki latar belakang bahasa Arab dengan pengalaman dan pemahaman mendalam terhadap bahasa tersebut, untuk memastikan bahwa setiap soal yang dibuat mencukupi standar yang baik. Keahlian pengajar dalam penyusunan soal tidak hanya mencakup kejelasan dan ketepatan, tetapi juga kesesuaian dengan kurikulum dan kebutuhan siswa [11], sehingga siswa dapat belajar dengan efektif dan mencapai hasil yang optimal dalam pembelajaran bahasa Arab. Ketika peneliti melakukan observasi awal, temuan yang didapatkan di lapangan bahwasannya soal penilaian tengah semester belum pernah dianalisis sebelumnya oleh pengajar bahasa arab MA Islamiyah, alasannya soal penilaian tersebut tidak digunakan kembali. Dengan demikian, peneliti perlu melakukan analisis secara kuantitatif terhadap butir soal penilaian untuk mengetahui kualitas setiap soal tersebut.

Dalam penyusunan tiap butir soal yang baik, harus mempertimbangkan beberapa kriteria, 1) aspek validitas, yaitu ukuran yang menunjukkan seberapa akurat suatu instrumen dapat mengukur apa yang seharusnya diukur. 2) reliabilitas soal, artinya suatu instrumen atau soal dapat menunjukkan seberapa konsisten hasilnya jika tes digunakan berkali-kali. 3) tingkat kesulitan soal adalah angka yang menunjukkan apakah butir soal tersebut berkategori sukar atau mudah. soal dikatakan baik jika pertanyaan tidak terlalu mudah atau terlalu sukar. Itulah sebabnya, suatu soal harus memiliki tingkat kesukaran yang seimbang. 4) daya pembeda soal, merupakan kemampuan suatu butir soal untuk membedakan antara siswa yang memiliki kemampuan tinggi dan yang memiliki kemampuan rendah. Butir soal dikatakan baik jika hanya siswa yang berkemampuan tinggi dapat menjawab pertanyaan dengan benar. 5) kualitas pengecoh (*distractor*), digunakan untuk melihat apakah option atau pilihan jawaban dalam soal pilihan ganda berfungsi dengan baik atau tidak. Pilihan jawaban yang baik akan digunakan kembali, dan yang tidak berfungsi perlu untuk diganti atau dirubah [12].

Software anates, yang dikembangkan oleh Karnoto dan Yudi Wibisono, dapat menganalisis soal pilihan ganda dan uraian secara otomatis. Program ini memungkinkan pendidik dapat menganalisis soal tes dengan cara menghitung skor, reliabilitas tes, tingkat kesulitan soal, daya pembeda, menghitung korelasi antara hasil pada nilai tiap soal dan hasil keseluruhan, juga dengan mengelompokkan nilai yang lebih tinggi atau lebih rendah, dan mengevaluasi efektivitas option/pilihan jawaban. Software anates memiliki beberapa kelebihan, seperti pemrograman yang sudah berbahasa indonesia, tampilan yang sederhana dan memberikan kemudahan bagi penggunaanya, perintah program yang jelas dan mudah diikuti, dan tentunya dapat memverifikasi jawaban benar dan salah pada tiap butir tes yang dibuat pendidik dengan praktis, cepat dan akurat. Selain itu, software anates ini memiliki kelemahan, yaitu pengisian data harus dilakukan secara manual dan hasil akhir akan berkurang jika terjadi kesalahan ketika memasukkan data. Oleh karenanya, dalam pengisian data pada software anates ini diperlukan ketelitian dan ketekunan [13].

Penelitian terkait analisis butir soal dibahas oleh peneliti lainnya. Peneliti menemukan tulisan, artikel, jurnal terkait analisis butir soal. Alfa Izzul Haq pada tahun 2024 dengan judul “Analisis Butir Soal Bahasa Arab kelas X SMA Muhammadiyah 2 Sidoarjo Berbasis Aplikasi Anates”. Adapun hasilnya menunjukkan bahwa soal memiliki validitas 80%, yang bermakna (signifikan), reliabilitasnya 0,87, taraf kesukarannya sedang, dan daya pembedanya 80%, atau

sebanyak 16 soal, dapat membedakan siswa yang memiliki pemahaman materi rendah dan tinggi [14]. Pada topik lain yang ditulis oleh Raudhatul Husna pada tahun 2022 dengan judul “Tahlili Bunuudi As’ilah Al-lughoh Al-arabiyyah fii Madrasah Aliyah bi Ma’had Jeumala Amal Pidi Jaya”. Diketahui bahwa hasil analisis validitas butir soal memperoleh (56%), reliabilitas memperoleh (0,90%), tingkat kesulitan soal berkategori mudah (36%), sedang (32%), dan sukar (32%), sedangkan hasil analisis daya pembeda soal memiliki daya pembeda yang cukup baik (32%) [15]. Kemudian dalam artikel yang berjudul “Analisis Soal Tematik Bahasa Arab berbasis HOTS pada Kelas XII MAN 3 Jember” ditulis oleh Fina Aunul Kafi dan Zidni Ilma pada tahun 2022. Penulis menguraikan hasil penelitiannya menggunakan anates versi 4.0.2 bahwa dari lima belas jumlah soal yang diuji, sembilan diantaranya belum memenuhi kriteria HOTS, validitasnya memperoleh 0,63, reliabilitas butir soal mencapai 0,78, dan tidak adanya soal yang tergolong sulit maupun sangat sulit, begitu juga kualitas alternatif jawaban nomor empat, sepuluh, sebelas, dan empat belas tidak memadai. Dari empat soal tersebut dinyatakan tidak layak, sedangkan yang lainnya memerlukan perbaikan. Maka dari itu, soal-soal tersebut masih membutuhkan revisi dari segi kuantitatif dan kualitatif [16].

Setiap penelitian pasti memiliki persamaan dan perbedaan tersendiri, termasuk objek penelitian, alat yang digunakan, dan jenis penelitian yang dimaksud. Penelitian sebelumnya dengan peneliti memiliki persamaan, yaitu sama-sama fokus menggunakan software anates, serta subjek dan objek yang dimaksud berbeda. Selain itu, yang membedakan penelitian ini dengan beberapa penelitian yang telah disebutkan sebelumnya yaitu dengan menambahkan analisis kualitas pengecoh soal atau berfungsi dan tidaknya distraktor, yang akan menjadikan pelengkap serta penyempurna pada penelitian sebelumnya. Maka, tujuan dari penelitian ini untuk mengetahui kualitas butir soal Bahasa arab siswa kelas XI MA Islamiyah Sidoarjo secara validitas, reliabilitas, daya pembeda, tingkat kesulitan dan efektivitas pengecoh soal dengan bantuan software anates. Dengan penelitian ini, peneliti berharap dapat memberikan pemahaman yang mendalam tentang kualitas butir soal dan memberikan rekomendasi pemikiran yang berharga untuk meningkatkan kualitas soal bahasa arab di MA Islamiyah Sidoarjo.

II. METODE

Penelitian ini dilakukan dengan menggunakan metode deskriptif kuantitatif. Penelitian deskriptif itu sendiri bertujuan untuk menggambarkan fenomena atau kejadian yang akan diteliti secara menyeluruh dan apa adanya [17]. Selanjutnya penelitian kuantitatif adalah jenis penelitian yang melibatkan angka-angka dan dianalisis dengan statistik [18]. Dalam penelitian kuantitatif deskriptif, peneliti merencanakan, melakukan, mengumpulkan, menganalisis, dan melaporkan hasil dari penelitian [19]. Penelitian ini dilaksanakan di MA Islamiyah Sidoarjo yang beralamat di Jalan Sumorame RT 01/ RW 04, Kecamatan Candi, Kabupaten Sidoarjo. Peneliti memilih sekolah ini dikarenakan berkaitan dengan tujuan penelitian. Adapun subjek penelitian terdiri dari 33 siswa yang diambil dari kelas XI IPA MA Islamiyah Sidoarjo pada tahun ajaran 2024/2025. Sementara itu, objek penelitian meliputi jawaban soal pilihan ganda siswa, mata pelajaran bahasa arab pada sumatif tengah semester ganjil. Teknik pengumpulan data yang digunakan peneliti adalah dokumentasi berupa soal penilaian, kunci jawaban, serta lembar jawaban hasil pekerjaan siswa. Hasil penelitian dianalisis dengan menggunakan bantuan software Anates versi 4.0.9 untuk menilai validitas, reliabilitas, tingkat kesulitan soal, daya pembeda serta efektivitas pengecoh.

III. HASIL DAN PEMBAHASAN

Melalui analisis yang dilakukan terhadap soal pilihan ganda pelajaran Bahasa Arab di MA Islamiyah Sidoarjo dengan menggunakan software Anates 4.0.9, telah didapatkan informasi mengenai validitas, reliabilitas, tingkat kesulitan, daya pembeda soal, dan kualitas pengecoh tiap item soal.

A. Validitas

Makna dari validitas merupakan sejauh mana instrumen yang dibuat oleh guru mampu mengukur aspek yang ingin diukur, atau dengan kata lain, ketepatan instrumen. Untuk menilai setiap butir soal, skor butir dapat dibandingkan dengan skor total. Teknik korelasi yang dapat digunakan untuk menganalisis validitas butir soal ini adalah korelasi product-moment. Apabila skor item menunjukkan kesesuaian atau keselarasan dengan skor total, maka item tersebut dapat dianggap tinggi atau dinyatakan valid. Dalam istilah statistik, terdapat korelasi positif yang signifikan antara skor butir soal dan skor total [20]. Berikut hasil analisis validitas menggunakan software anates:

```
KORELASI SKOR BUTIR DG SKOR TOTAL
=====
Jumlah Subyek= 33
Butir Soal= 10
Nama berkas: C:\USERS\YUMNA HANIFA\ONEDRIVE\DOCUMENTS\TUGAS AKHIR\DATA JAWABAN.ANA

No Butir Baru  No Butir Asli      Korelasi   Signifikansi
1              1                 0,749     Sangat Signifikan
2              2                 0,379     -
3              3                 0,853     Sangat Signifikan
4              4                 0,545     -
5              5                 0,793     Sangat Signifikan
6              6                 0,620     Signifikan
7              7                 0,284     -
8              8                 0,883     Sangat Signifikan
9              9                 0,491     -
10             10                0,578     Signifikan
```

Gambar 2. Hasil analisis validitas

Catatan: Batas signifikansi koefisien korelasi sebagai berikut:

df (N-2)	P=0,05	P=0,01	df (N-2)	P=0,05	P=0,01
10	0,576	0,708	60	0,250	0,325
15	0,482	0,606	70	0,233	0,302
20	0,423	0,549	80	0,217	0,283
25	0,381	0,496	90	0,205	0,267
30	0,349	0,449	100	0,195	0,254
40	0,304	0,393	125	0,174	0,228
50	0,273	0,354	>150	0,159	0,208

Bila koefisien = 0,000 berarti tidak dapat dihitung.

Gambar 1. Batas signifikansi koefisien kolerasi

Menghitung koefisien korelasi antara skor tiap butir soal dan skor total adalah cara untuk mengetahui apakah setiap butir soal dalam tes homogen. Perhitungan ini dilakukan untuk setiap butir soal dalam tes tersebut. Sebuah butir soal dianggap homogen jika koefisien kolerasinya sama atau diatas batas signifikansi. Sebaliknya, apabila koefisien kolerasinya negatif atau lebih rendah dari batas signifikansi, maka item tersebut dianggap tidak homogen [13]. Pada gambar 2 diketahui 10 butir soal Bahasa arab, batas signifikansi koefisien kolerasinya berada diantara 0,576 sampai 0,708 yang berarti hasilnya signifikan. Jika nilainya lebih dari 0,708, maka hasilnya sangat signifikan. Dengan demikian, jika nilai kolerasinya lebih rendah dari 0,576, maka tes tersebut tidak valid atau tidak signifikan.

Hasil analisis validitas pada gambar 1 menunjukkan bahwa dari 10 butir soal pilihan ganda Bahasa arab menunjukkan bahwa terdapat 60% tergolong valid karena kolerasinya diatas 0,576, yaitu nomor 1, 3, 5, 6, 8, dan 10. Sementara itu, 40% tergolong tidak valid sebab tidak ditemukan nilai kolerasinya, seperti nomor 2, 4, 7, 9. Butir soal yang tidak homogen atau tidak valid ini cenderung kurang efektif dalam mengukur kemampuan siswa yang seharusnya diukur oleh guru. Dari keempat butir soal pilihan ganda yang tidak valid, terdapat soal nomor 2 dan 4 yang berhubungan dengan menerjemahkan kosakata, nomor 7 yang berhubungan dengan menentukan ism makan, serta nomor 9 yang berhubungan tentang na'at man'ut atau sifat terkait dengan as-su'uq. Sebaiknya, empat butir soal ini harus dievaluasi kembali oleh guru atau dihapus karena tidak memiliki validitas.

B. Reliabilitas

Reliabilitas didefinisikan sebagai kemampuan suatu tes untuk mengukur apa yang seharusnya diukur sehingga menghasilkan informasi yang dapat dipercaya. Dengan kata lain, reliabilitas menggambarkan seberapa konsisten butir soal dapat memberikan hasil yang serupa ketika diuji coba berulang kali [21]. Arikunto juga menjelaskan bahwa suatu tes dapat memberikan hasil yang konsisten, maka tes tersebut dapat dianggap memiliki taraf kepercayaan yang tinggi. Jika hasilnya berubah-ubah, perubahan yang terjadi dapat dianggap tidak signifikan [12]. Terdapat tiga cara untuk menghitung reliabilitas tes, 1) Pendekatan bentuk paralel atau tes ekuivalen, 2) Pendekatan tes ulang, dan 3) Pendekatan belah dua. Dalam aplikasi anates ini, reliabilitas tes dihitung menggunakan pendekatan belah dua, yaitu dengan membagi item atau butir soal berdasarkan ganjil-genap dan membagi atas item-item awal-akhir [13]. Tabel berikut menunjukkan kriteria yang digunakan untuk menafsirkan reliabilitas setiap item penilaian.

Tabel 1. Interpretasi reliabilitas soal

Reliabilitas soal	Interpretasi
0,800 – 1,000	Sangat tinggi
0,600 – 0,799	Tinggi
0,400 – 0,599	Cukup
0,200 – 0,399	Rendah
0,00 – 0,199	Sangat rendah

Berikut adalah hasil analisis reliabilitas tes yang diperoleh dari software anates:

RELIABILITAS TES						
=====						
Rata2= 8,06						
Simpang Baku= 2,40						
KorelasiXY= 0,82						
Reliabilitas Tes= 0,90						
Nama berkas: C:\USERS\YUMNA HANIFA\ONEDRIVE\DOCUMENTS\TUGAS AKHIR\DATA JAWABAN.ANA						
No.Ucut	No.	Subyek	Kode/Nama Subyek	Skor Ganjil	Skor Genap	Skor Total
1	1	Abdul Rokhman	3	5	8	
2	2	Aditya Jaya P...	2	1	3	
3	3	Ahmad Raihan	4	5	9	
4	4	Ainur Rochimah	4	5	9	
5	5	Ainur Rochmah	4	5	9	
6	6	Alnoora Fajar ...	4	5	9	
7	7	Amelia Kartik...	2	3	5	
8	8	Arya Octo Man...	2	3	5	
9	9	Aprillya Nur ...	5	5	10	
10	10	Chelsea Budia...	4	5	9	
11	11	Durrutun Nafisah	5	5	10	
12	12	Fanny Dwi Azz...	1	2	3	
13	13	Fidyah Mila A...	4	5	9	
14	14	Fitria Hanifa	5	5	10	
15	15	Liatul Aini	1	3	4	
16	16	M. Syafiq Alm...	5	4	9	
17	17	Macelatus Soikha	5	5	10	
18	18	Marteen Julian	2	3	5	
19	19	Marwah	3	4	7	
20	20	Moch. Mashuri	2	3	5	
21	21	Muhammad Fachri	4	5	9	
22	22	Nafisah Rahma...	1	2	3	
23	23	Nur Azizah	5	5	10	
24	24	Nur Laily	5	5	10	
25	25	Qeisya Dwi Fe...	5	5	10	
26	26	Rasya Rahmah ...	5	4	9	
27	27	Rosalinda Dwi...	5	4	9	
28	28	Sofia Abdilla...	5	5	10	
29	29	Syintia Afif...	5	5	10	
30	30	Umi Soleha	4	5	9	
31	31	Vevi Aminatuz...	5	5	10	
32	32	Visca Auliya ...	4	5	9	
33	33	Yeni Nurfaridah	5	5	10	

Gambar 3. Hasil analisis reliabilitas tes

Analisis reliabilitas tes yang diperoleh melalui software Anates menunjukkan bahwa dari 10 butir soal Sumatif Tengah Semester (STS) Bahasa Arab tahun ajaran 2024/2025 nilai reliabilitasnya adalah 0,90. Nilai ini berada dalam rentang 0,800 hingga 1,000, yang mengindikasikan bahwa tes tersebut memiliki tingkat konsistensi yang sangat tinggi. Dengan demikian, skor yang dihasilkan dari tes ini dapat diandalkan atau dipercaya. Karena hasil analisis menunjukkan bahwa soal-soal yang diuji telah sesuai standar atau syarat yang telah ditetapkan. Dengan demikian, reliabilitas soal sumatif tengah semester Bahasa arab dianggap layak dan dapat digunakan dalam tes tanpa memerlukan perubahan atau perbaikan.

C. Taraf kesukaran soal

Penilaian kesukaran soal merupakan penilaian mengenai sejauh mana kesulitan suatu soal. Jika suatu soal memiliki tingkat kesukaran yang seimbang atau proporsional, maka soal tersebut dapat dianggap baik. Sebuah soal tes sebaiknya

tidak terlalu sulit maupun terlalu mudah. Azmawi Zainul juga menyatakan bahwa tingkat kesukaran butir soal hanya menunjukkan seberapa sulit atau mudah suatu soal bagi kelompok peserta tes tertentu, dan tidak mencerminkan apakah butir soal tersebut baik atau tidak. Penilaian hasil belajar yang terlalu rumit atau terlalu mudah tidak memberikan informasi yang cukup tentang peserta tes atau butir soal [13].

Tingkat kesukaran soal merujuk pada kemungkinan untuk menjawab soal dengan benar berdasarkan kemampuan tertentu, yang umumnya diukur dengan indeks. Indeks kesukaran atau *difficulty index* memiliki rentang nilai antara 0,00 hingga 1,0. Semakin tinggi indeksnya, semakin mudah soal tersebut, sementara semakin rendah indeksnya, semakin sulit soal tersebut [12]. Melihat hasil analisis kesukaran soal tersebut disajikan dalam bentuk persentase, maka diperlukan untuk mengubah indeks kesukaran ke bentuk desimal.

Tabel 2. Indeks kesukaran soal

Indeks kesukaran	Tafsiran
P 1,00 – 0,30	Soal sukar
P 0,30 – 0,70	Soal sedang
P 0,70 – 0,85	Soal mudah

Berikut hasil analisis tingkat kesukaran butir soal:

TINGKAT KESUKARAN						
=====						
Jumlah Subjek= 33						
Butir Soal= 10						
Nama berkas:	C:\USERS\YUMNA HANIFA\ONEDRIVE\DOCUMENTS\TUGAS AKHIR\DATA JAWABAN.ANA					
No	Butir Baru	No Butir Asli	Jml Betul	Tkt. Kesukaran(%)	Tafsiran	
1		1	19	57,58	Sedang	
2		2	32	96,97	Sangat Mudah	
3		3	25	75,76	Mudah	
4		4	31	93,94	Sangat Mudah	
5		5	25	75,76	Mudah	
6		6	28	84,85	Mudah	
7		7	25	75,76	Mudah	
8		8	23	69,70	Sedang	
9		9	31	93,94	Sangat Mudah	
10		10	27	81,82	Mudah	

Gambar 4. Hasil analisis kesukaran soal

Berdasarkan pada Gambar 4, hasil analisis tersebut disajikan dalam bentuk persentase. Untuk memudahkan pemahaman, diperlukan untuk mengubah indeks kesulitan ke dalam format desimal. Dalam gambar tersebut, terdapat 3 butir soal (nomor 2, 4, 9) yang berkategori sangat mudah, 5 butir soal (3, 5, 6, 7, 10) yang berkategori mudah, dan 2 butir soal yang berkategori sedang yakni nomor 1 dan 8. Sementara itu, tidak ditemukan soal yang berkategori sukar atau sangat sukar. Mengingat tingkat kesulitan soal ini digunakan untuk membedakan siswa dengan kemampuan tinggi, sedang, dan rendah, disarankan untuk mengganti soal yang berkategori sangat mudah [12]. Mengingat bahwa tingkat kesulitan soal tergolong rendah, yang mungkin perlu dievaluasi untuk memastikan variasi dalam tingkat kesulitan dalam penilaian. Secara keseluruhan, analisis dalam penelitian ini menunjukkan bahwa siswa dapat dengan mudah menjawab Sebagian besar pertanyaan, yang mungkin menandakan perlunya perubahan untuk meningkatkan kesulitan dan efektivitas penilaian agar dapat masuk ke dalam kategori sedang.

D. Daya pembeda butir soal

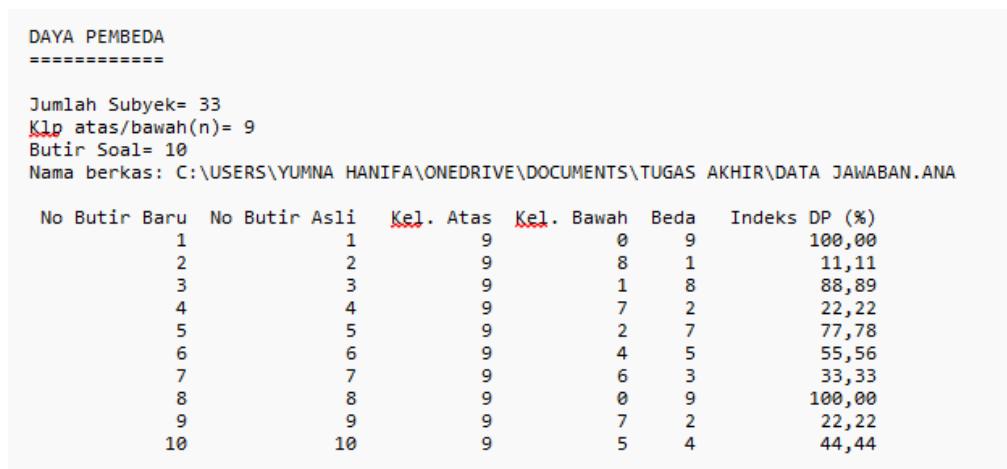
Daya pembeda soal merupakan indikator yang menggambarkan seberapa efektif butir soal dalam membedakan antara kelompok siswa yang unggul dan kelompok siswa asor atau yang lebih rendah [22]. Untuk mengetahui daya pembeda soal, dapat dilihat dari nilai indeks diskriminasi yang dimiliki soal tersebut. Soal dengan indeks yang lebih tinggi lebih efektif dalam membedakan siswa yang lebih pandai dengan siswa yang kurang pandai. Manfaat dari daya pembeda soal ini adalah untuk meningkatkan kualitas setiap butir soal serta untuk mengetahui sejauh mana kemampuan siswa dapat diidentifikasi atau dibedakan. [13]. Sehingga, daya pembeda ini dapat menentukan bahwa semakin tinggi nilai daya pembedanya, semakin baik kemampuan soal tersebut dalam membedakan antara kelompok unggul dan kelompok bawah dengan sangat baik. Pada tabel berikut menunjukkan interpretasi daya pembeda butir soal.

Daya pembeda soal adalah indikator yang menunjukkan seberapa efektif suatu butir soal dalam membedakan antara kelompok unggul (siswa yang menguasai materi dengan baik) dan kelompok rendah (siswa yang kurang menguasai materi). Untuk menilai daya pembeda, dapat dilihat dari besar atau kecilnya indeks diskriminasi. Soal dengan indeks yang lebih tinggi lebih mampu membedakan siswa yang pandai dari siswa yang kurang pandai. Manfaat dari daya pembeda ini adalah untuk meningkatkan kualitas setiap butir soal dan untuk mengetahui sejauh mana kemampuan siswa dapat diidentifikasi atau dibedakan. Dengan demikian, semakin tinggi nilai daya pembedanya, semakin baik kemampuan soal tersebut dalam membedakan antara kelompok unggul dan kelompok rendah. Tabel berikut menyajikan interpretasi daya pembeda butir soal.

Tabel 3. Indeks daya pembeda soal

Indeks diskriminasi	Interpretasi
0,00 – 0,20	Tidak baik
0,20 – 0,40	Cukup
0,40 – 0,70	Baik
0,70 – 0,10	Baik sekali

Berikut hasil uji daya pembeda tiap butir soal yang diperoleh dari software anates:



Gambar 5. Hasil uji daya pembeda soal

Berdasarkan pada gambar 5 menunjukkan bahwa hasil uji daya pembeda dari 10 soal terdapat sebanyak 40% butir soal yang termasuk dalam kategori baik sekali (nomor 1, 3, 5, 8), sebanyak 20% butir soal termasuk dalam kategori baik (nomor 6, 10), dan sebanyak 30% butir soal termasuk kategori cukup (nomor 4, 7, 9), serta 10% butir soal sisanya berkategori jelek (nomor 2). Rendahnya daya pembeda pada soal nomor 2 disebabkan oleh banyaknya jawaban benar dari kedua kelompok, sehingga soal tersebut dianggap kurang baik. Soal dengan daya pembeda yang sangat baik, baik, dan cukup dianggap memenuhi kriteria dan perlu dipertahankan. Di sisi lain, soal dengan daya pembeda rendah atau kurang baik perlu diperbaiki secara menyeluruh dengan memahami alasan kegagalan soal tersebut. Salah satu langkah yang dapat diambil adalah memperjelas soal yang tidak jelas agar tidak menimbulkan kebingungan bagi siswa yang memiliki kemampuan tinggi dalam menjawab [21].

E. Kualitas pengecoh soal

Pengecoh atau distraktor merupakan elemen penting dalam soal pilihan ganda. Sebagai alternatif jawaban yang salah/menyesatkan, distraktor dirancang untuk menarik perhatian peserta ujian yang kurang memahami materi. Fungsi utama pengecoh adalah membedakan peserta yang memiliki pemahaman baik dari mereka yang kurang memahami konsep yang diujikan. Dengan demikian, pengecoh yang efektif akan membantu memastikan bahwa soal memiliki daya diskriminasi yang baik. Selain itu, pengecoh yang berkualitas berkontribusi pada validitas soal secara keseluruhan, karena mereka menguji kemampuan peserta untuk memilih jawaban yang benar di antara pilihan-pilihan yang logis dan relevan [25].

Kriteria untuk menentukan efektivitas pengecoh telah dirumuskan dengan mempertimbangkan distribusi jawaban peserta ujian. Sebuah pengecoh dianggap efektif jika dipilih oleh setidaknya 5% dari total peserta ujian. Dalam konteks penelitian ini, dimana jumlah responden ialah 33 orang, maka minimal dua orang harus memilih pengecoh agar dapat

dikategorikan sebagai efektif. Jika pengecoh tidak memenuhi kriteria ini, maka ia dianggap tidak menarik atau tidak sesuai dengan konteks soal, sehingga perlu direvisi [26].

Kualitas Pengecoh								
Jml Subyek= 33		Butir Soal = 10 ** : 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	3++	10--	0--	1-	19**	0	
2	2	0--	32**	1--	0--	0--	0	
3	3	25**	1-	2++	1-	4--	0	
4	4	0--	2--	0--	31**	0--	0	
5	5	0--	25**	0--	1-	7---	0	
6	6	1++	28**	0--	0--	4--	0	
7	7	1-	25**	0--	5--	2++	0	
8	8	1-	2++	6---	1-	23**	0	
9	9	31**	1--	1--	0--	0--	0	
10	10	27**	1+	0--	3--	2+	0	

Gambar 6. Hasil analisis efektivitas pengecoh

Berdasarkan hasil analisis yang ditampilkan pada gambar 6, tidak semua pengecoh dalam soal-soal yang dianalisis memenuhi kriteria efektivitas. Beberapa pengecoh terlihat berhasil menarik perhatian peserta ujian sesuai dengan kriteria minimal, sementara yang lainnya sama sekali tidak dipilih atau hanya menarik perhatian satu responden saja. Sebagai contoh, pada soal nomor 1, distraktor a dipilih oleh tiga responden, sehingga memenuhi kriteria sebagai pengecoh yang baik. Distraktor ini berhasil menarik perhatian peserta yang kurang memahami materi, tanpa terlalu mencolok sebagai jawaban yang salah. Sebaliknya, distraktor b, meskipun dipilih oleh sepuluh responden, dinilai kurang efektif. Alasannya karena pengecoh ini mungkin terlalu mudah dikenali sebagai jawaban yang salah, sehingga tidak dapat membedakan dengan baik antara peserta yang memiliki pemahaman rendah dan mereka yang memiliki pemahaman sedang. Hal ini mengindikasikan bahwa meskipun sebuah pengecoh dipilih oleh banyak peserta, kualitasnya tetap perlu dievaluasi berdasarkan daya tarik logisnya dalam soal.

Contoh lainnya dapat dilihat pada soal nomor 3. Distraktor a sangat efektif, memiliki 25 responden. Hal ini menunjukkan bahwa pengecoh tersebut dirancang dengan baik dan mampu menarik perhatian banyak peserta yang kurang memahami materi. Namun, pengecoh lainnya, seperti pilihan b, pilihan c, dan pilihan d, kurang efektif karena masing-masing hanya dipilih oleh satu responden atau tidak dipilih sama sekali. Distraktor semacam ini perlu direvisi agar dapat meningkatkan daya tariknya dan berkontribusi lebih baik terhadap kualitas soal. Untuk pengecoh yang sudah memenuhi kriteria minimal, yaitu yang berwarna hijau atau memiliki simbol + atau ++ dalam hasil analisis, tidak diperlukannya banyak perubahan. Pengecoh seperti ini berhasil menjalankan fungsinya dengan baik dan cukup dipertahankan apa adanya. Namun pengecoh yang tidak dipilih sama sekali (nilai 0) atau hanya dipilih oleh satu responden, maka diperlukan perbaikan.

Tabel 4. Ringkasan keberfungsiannya option jawaban

Kriteria	Butir soal	Total soal	Percentase (%)
Berfungsi	2, 3, 6, 7, 8, 10	7	70%
Tidak Berfungsi	1, 4, 9	3	30%

Berdasarkan analisis kualitas pengecoh pada 10 butir soal pilihan ganda Bahasa Arab, terdapat variasi dalam efektivitas pengecoh pada setiap soal. Sebanyak 70% dari soal nomor 2, 3, 5, 6, 7, 8, dan 10 memiliki pengecoh yang berfungsi dengan baik, alasannya karena Sebagian besar opsi pilihan dapat menarik perhatian peserta ujian. Soal nomor 6 dan 7 terutama memiliki pengecoh yang sangat efektif, karena hampir semua opsi dirancang untuk menghindari kesalahan peserta. Ini menunjukkan bahwa soal-soal tersebut memiliki tingkat kesulitan yang sesuai dan dapat membedakan kemampuan peserta. Sebaliknya, pengecoh yang sebagian besar tidak berfungsi ditemukan dalam persentase 30%, seperti soal nomor 1, 4, dan 9. Diperlukannya revisi untuk meningkatkan kualitas pengecoh karena opsi dalam soal-soal ini tidak mampu menarik perhatian peserta secara signifikan.

Pengecoh yang tidak efektif harus dirancang ulang agar lebih menarik dan relevan. Perbaikan dapat dilakukan dengan memastikan bahwa pengecoh menyerupai jawaban yang benar atau mencerminkan kesalahan umum yang mungkin dilakukan peserta. Setelah direvisi, soal harus diuji ulang untuk memastikan bahwa perubahan telah menghasilkan hasil yang diinginkan [23]. Dengan tindakan ini, kualitas soal pilihan ganda dapat terus ditingkatkan, sehingga peserta dapat dibedakan berdasarkan tingkat pemahaman mereka. Pengecoh yang dirancang dengan baik tidak hanya memastikan bahwa soal valid, tetapi juga memastikan bahwa soal tersebut adil dan secara akurat menunjukkan kemampuan peserta. Oleh karena itu, pengecoh yang efektif sangat penting untuk mendukung tujuan evaluasi yang diinginkan dalam soal pilihan ganda.

IV. SIMPULAN

Simpulan Simpulan yang diperoleh dari analisis butir soal pilihan ganda bahasa arab siswa kelas XI MA Islamiyah Sidoarjo, didapatkan dari segi validitasnya yakni terdapat 6 butir soal yang dianggap valid karena kolerasinya diatas 0,576, sedangkan 4 butir soal dianggap tidak valid karena tidak ditemukan nilai kolerasinya. Segi reliabilitasnya menunjukkan hasil bahwa reliabilitas soal tersebut sangat tinggi dengan nilai 0,90. Dari segi tingkat kesukaran soalnya menunjukkan bahwa tidak ditemukannya butir soal yang berkategori sukar atau sangat sukar, hanya terdapat 2 butir soal berkategori sedang, 3 butir soal berkategori sangat mudah, dan 5 butir soal berkategori mudah. Dari segi daya pembedanya yaitu terdapat 4 butir soal mempunyai daya beda yang baik sekali, 2 butir soal mempunyai daya beda yang baik, 3 butir soal mempunyai daya beda yang cukup dan 1 butir soal mempunyai daya beda yang jelek atau perlu direvisi. Selanjutnya dari segi kualitas alternatif jawaban dari 10 butir soal terdapat 7 butir soal yang berfungsi dengan baik dan 3 butir soal tidak berfungsi sebagai distraktor yang baik.

REFERENSI

- [1] U. Ridho, "Evaluasi Dalam Pembelajaran Bahasa Arab," *An Nabighoh J. Pendidik. dan Pembelajaran Bhs. Arab*, vol. 20, no. 01, p. 19, 2018, doi: 10.32332/an-nabighoh.v20i01.1124.
- [2] Y. Monica, S., dan Sudarman, "ANALISIS BUTIR SOAL UJIAN TENGAH SEMESTER GANJIL SENI BUDAYA KELAS VII DI SMPN 29 SIJUNJUNG," *J. Chem. Inf. Model.* 53(9)1689–99, vol. 8, no. 3, pp. 1–8, 2019, doi: <https://doi.org/10.24036/jsu.v7i3.103276>.
- [3] E. N. Taib, "Analisis Kualitas Aspek Materi Butir Soal Buatan Dosen," *Biot. J. Ilm. Biol. Teknol. dan Kependidikan*, vol. 2, no. 2, pp. 77–137, 2014, doi: 10.22373/biotik.v2i2.245.
- [4] I. Magdalena, S. N. Fauziah, S. N. Faziah, and F. S. Nupus, "Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas III SDN Karet 1 Sepatan," *BINTANG J. Pendidik. dan Sains*, vol. 3, no. 2, pp. 198–214, 2021.
- [5] I. Magdalena, G. Ramadhan, H. D. Wahyuni, and N. D. Safitri, "Pentingnya Proses Evaluasi Dalam Pembelajaran Di Sekolah Dasar," *Ta'rim J. Pendidik. dan Anak Usia Dini*, vol. 4, no. 3, pp. 167–176, 2023, doi: 10.59059/tarim.v4i3.220.
- [6] R. A. Tho'imah, M. H. Royyan, and A. A. A. Ghani, *Ta'lim al-Arabiyyah lighairi al-Naathiqa Biha; Manahijuhi wa Asalibuhu*. Jakarta: Raja Grafindo Persada, 1989.
- [7] Mulyadi, *Evaluasi Pendidikan Pengembangan Model Evaluasi Pendidikan Agama Di Sekolah*. Malang: UIN-Maliki Press, 2014.
- [8] A. S. Ma'arif, "Sighah Ikhtibarat al-'Arabiyyah fi Dui al-Ikhtibar al-Mutqn (TOAFL)," *J. Al Bayan J. Jur. Pendidik. Bhs. Arab*, vol. 9, no. 2, pp. 160–173, 2018.
- [9] B. Fahmi, S. Rizqi, and N. E. Harmelinda, "Analisis Butir Soal Bahasa Arab MAS Pondok Pesantren Assalam Kampar Riau," *Ta'lim al-'Arabiyyah J. Pendidik. Bhs. Arab Kebahasaaraban*, vol. 6, no. 1, pp. 95–105, 2022, doi: 10.15575/jpba.v6i1.16193.
- [10] I. Rahmi, Syihabudin, and A. Sopian, "ANALISIS KUALITAS TES BAHASA ARAB BERBASIS HIGHER ORDER THINKING SKILL (HOTS)," *LISANUNA J. Ilmu Bhs. Arab dan Pembelajarannya*, vol. 10, no. 1, p. 45, 2020, doi: 10.22373/lis.v10i1.7805.
- [11] R. D. Nawanti, S. Purnomo, W. T. Santoso, A. Muhibbin, and S. Sumardi, "Peran Guru dalam Evaluasi Pembelajaran di Kelas," *JIIP - J. Ilm. Ilmu Pendidik.*, vol. 7, no. 5, pp. 4759–4765, 2024, doi: 10.54371/jiip.v7i5.4288.
- [12] S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan*, 3rd ed. Jakarta: Bumi Aksara, 2018.
- [13] S. Wiguna, *Aplikasi Anates Dalam Evaluasi Pembelajaran*. Banyumas: CV. Pena Persada, 2021.
- [14] A. I. Haq and N. Anwar, "Analisis Butir Soal Bahasa Arab kelas X SMA Muhammadiyah 2 Sidoarjo Berbasis Aplikasi Anates," vol. 12, no. 2, pp. 1–13, 2024, doi: <https://doi.org/10.21070/ijis.v12i2.1713>.
- [15] R. Husna, "Tahlili Bunuudi As'ilah Al-lughoh Al-arabiyyah fii Madrasah Aliyah bi Ma'had Jeumala Amal Pidi Jaya," *UIN Ar-raniry*, 2022.

- [16] F. A. Kafi and Z. Ilma, “Analisis Soal Tematik Bahasa Arab berbasis HOTS pada Kelas XII MAN 3 Jember,” *Al-Fusha Arab. Lang. Educ. J.*, vol. 4, no. 1, pp. 1–9, 2022, doi: <https://doi.org/10.36835/alfusha.v4i1.729>.
- [17] R. Agustianti *et al.*, *Metode Penelitian Kuantitatif Dan Kualitatif*. Tohar Media, 2022.
- [18] M. Ronaldy Aji Saputra, F. Idham Chalid, and H. Budianto, *Metode Ilmiah dan Penelitian: Kuantitatif, Kualitatif, dan Kepustakaan (Bahan Ajar Madrasah Riset)*. Sidoarjo: Nizamia Learning Center, 2023.
- [19] Sugiyono, *Metode Penelitian Kuantitatif*. Bandung: Alfabeta, 2022.
- [20] A. Munip, *Penilaian Pembelajaran Bahasa Arab*, no. May. Yogyakarta: Fakultas Tarbiyah dan Ilmu Keguruan (FITK) UIN Sunan Kalijaga, 2019.
- [21] W. Warju, S. R. Ariyanto, S. Soeryanto, and R. A. Trisna, “Analisis Kualitas Butir Soal Tipe Hots Pada Kompetensi Sistem Rem Di Sekolah Menengah Kejuruan,” *J. Pendidik. Teknol. dan Kejuru.*, vol. 17, no. 1, p. 95, 2020, doi: 10.23887/jptk-undiksha.v17i1.22914.
- [22] W. Buska and R. Widayanti, “Taqniyyah Al Ikhtibārah Fī Al-Ta’Līmi Al-Lughah Al-‘Arabiyyah Bil-Jāmi’ah Al Islāmiyyah,” *Nazharat J. Kebud.*, vol. 26, no. 02, pp. 457–480, 2020, doi: 10.30631/nazharat.v26i02.38.
- [23] D. Putri Andini and M. Mukhlis, “Analisis Butir Soal pada Instrumen Asesmen Kompetensi Minimum Literasi Membaca di SMP IT Insan Utama Pekanbaru,” *Diglosia J. Kaji. Bahasa, Sastra, dan Pengajarannya*, vol. 6, no. 2, pp. 401–412, 2023, doi: 10.30872/diglosia.v6i2.658.

Conflict of Interest Statement:

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.