Shaut Al-'Arabiyah
P-ISSN : 2354-564X; E-ISSN : 2550-0317
Vol. 13 No. 2, Desember 2025
DOI : https://doi.org/10.24252/saa.v13i2.62594

Analisis Tes Pilihan Ganda Mata Pelajaran Bahasa Arab Mts Al-Hidayah Kota Batu

Nasrudin[1], Ela Saleha[2], Afriana Santosa[3] & Nur Hasanah[4]
[1,2,3,4]Pendidikan Bahasa Arab, Sekolah Tinggi Agama Islam Jarinabi
Corresponding E-mail: nasrudin98756@gmail.com

Abstrak: Penelitian ini bertujuan untuk menganalisis tingkat kesukaran dan daya pembeda butir soal pilihan ganda pada mata pelajaran Bahasa Arab di MTs Al-Hidayah Kota Batu. Analisis butir soal dilakukan untuk mengetahui sejauh mana setiap butir mampu mengukur kemampuan siswa secara proporsional dan membedakan antara peserta didik berkemampuan tinggi dan rendah. Metode penelitian yang digunakan adalah pendekatan kuantitatif deskriptif dengan jumlah sampel sebanyak 25 butir soal yang diujikan kepada seluruh siswa kelas VIII. Data dianalisis menggunakan indeks kesukaran (P) dan indeks daya pembeda (D) dengan kriteria interpretasi standar pengukuran klasik. Hasil penelitian menunjukkan bahwa 64% butir soal tergolong mudah, 32% sedang, dan 4% sukar, yang menandakan komposisi soal belum proporsional. Analisis daya pembeda menunjukkan 56% soal memiliki daya pembeda baik hingga sangat baik, sementara 24% tergolong sangat rendah dengan enam butir soal bernilai diskriminasi negatif. Hubungan antara tingkat kesukaran dan daya pembeda menunjukkan pola yang konsisten, di mana soal dengan tingkat kesukaran sedang memiliki daya pembeda lebih optimal. Secara keseluruhan, terdapat 5 butir soal dengan kategori sangat baik, 11 butir baik dengan perbaikan minor, 2 butir perlu revisi substansi, dan 7 butir perlu revisi total. Hasil penelitian ini menegaskan pentingnya analisis butir soal secara empiris sebagai dasar penyusunan instrumen evaluasi yang valid dan reliabel, sehingga dapat meningkatkan kualitas penilaian hasil belajar Bahasa Arab di madrasah.

Kata Kunci: Analisis Butir Soal; Tingkat Kesukaran; Daya Pembeda; Bahasa Arab; Evaluasi Pembelajaran

Abstract: *This study aims to analyze the level of difficulty and discrimination power of multiple-choice questions in Arabic language subjects at MTs Al-Hidayah in Batu City. The analysis of the questions was conducted to determine the extent to which each question was able to measure students' abilities proportionally and distinguish between high and low ability students. The research method used was a descriptive quantitative approach with a sample size of 25 questions administered to all eighth-grade students. The data were analyzed using the difficulty index (P) and discrimination index (D) with standard classical measurement interpretation criteria. The results showed that 64% of the items were easy, 32% were moderate, and 4% were difficult, indicating that the composition of the items was not yet proportional. The discrimination analysis showed that 56% of the items had good to very good discrimination, while 24% were classified as very low with six items having negative discrimination. The relationship between difficulty level and discriminating power shows a consistent pattern, whereby questions with moderate difficulty levels have optimal discriminating power. Overall, there were 5 items categorized as excellent, 11 items categorized as good with minor improvements, 2 items requiring substantial revision, and 7 items requiring total revision. The results of this study emphasize the importance of empirical item analysis as the basis for developing valid and reliable evaluation instruments, thereby improving the quality of Arabic language learning assessment in madrasahs.*

Keywords: *Item Analysis; Difficulty Level; Discrimination Power; Arabic Language; Learning Evaluation*

PENDAHULUAN

Peningkatan kualitas pendidikan nasional tidak dapat dilepaskan dari proses evaluasi pembelajaran yang terencana dan sistematis. Evaluasi berfungsi sebagai alat kontrol mutu pendidikan yang memberikan informasi akurat mengenai capaian belajar peserta didik. Hal ini ditegaskan dalam Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional Pasal 57 ayat (1), bahwa evaluasi merupakan bagian dari upaya pengendalian mutu pendidikan secara nasional sebagai bentuk akuntabilitas penyelenggaraan pendidikan terhadap seluruh pemangku kepentingan (Hidayat, T., Rizal, A. S., & Fahrudin, 2018). Evaluasi bukan hanya sekadar menilai hasil belajar, tetapi juga sebagai dasar untuk memperbaiki strategi pembelajaran dan pengembangan kurikulum agar lebih efektif dan adaptif terhadap kebutuhan peserta didik (Faishal, 2017).

Dalam praktiknya, evaluasi pembelajaran tidak dapat dipisahkan dari kegiatan pengukuran dan penilaian. Pengukuran berperan untuk mengumpulkan data tentang kemampuan peserta didik, baik dalam bentuk kuantitatif seperti skor atau nilai tes, maupun kualitatif berupa deskripsi kemampuan dan sikap siswa (Mar & Hilmi, 2021). Dengan demikian, pengukuran menjadi fondasi bagi proses evaluasi yang objektif dan valid (Nadya et al., 2024). Dalam konteks pengajaran, terutama pengajaran bahasa, evaluasi memiliki posisi yang sangat penting karena membantu guru memahami sejauh mana tujuan pembelajaran telah tercapai. Sebagaimana dijelaskan oleh (Maulana & Sanusi, 2020), pengajaran terdiri dari tiga komponen yang saling terkait: tujuan pembelajaran, pelaksanaan pembelajaran, dan evaluasi hasil pembelajaran. Komponen evaluasi memastikan bahwa dua komponen sebelumnya dapat berjalan efektif dan menghasilkan peningkatan kualitas belajar.

Salah satu instrumen utama dalam evaluasi pendidikan adalah tes hasil belajar. Tes digunakan untuk mengukur sejauh mana peserta didik memahami materi pelajaran, serta untuk menilai perkembangan dan kemajuan belajar. Berdasarkan bentuknya, tes dapat dibedakan menjadi tes objektif (seperti pilihan ganda) dan tes subjektif (seperti uraian). Tes objektif, khususnya bentuk pilihan ganda (multiple choice), dianggap paling efisien dalam mengukur pengetahuan, pemahaman, dan kemampuan berpikir tingkat tinggi atau High Order Thinking Skills (HOTS) secara luas (Ramadani & Handayani, 2024). Tes ini memberikan kemudahan dalam penilaian yang cepat, objektif, dan konsisten, sekaligus memungkinkan pengukuran terhadap berbagai level kemampuan kognitif siswa (Oktaviani, 2017).

Namun demikian, permasalahan utama yang masih sering terjadi di lapangan adalah bahwa banyak guru belum memperhatikan prinsip-prinsip dasar penyusunan tes yang baik. Dalam penyusunan instrumen tes, guru seringkali hanya mengandalkan soal yang tersedia di buku teks atau sumber lain tanpa memastikan kesesuaian dengan Kompetensi Inti (KI) dan Kompetensi Dasar (KD) dalam kurikulum yang berlaku. Akibatnya, butir soal yang digunakan tidak selalu selaras dengan tujuan pembelajaran yang ingin dicapai dan berpotensi menghasilkan penilaian yang tidak akurat terhadap kemampuan peserta didik (Somahhida et al., 2022). Selain itu, masih banyak guru yang tidak melakukan analisis validitas, reliabilitas, tingkat kesukaran, dan daya pembeda terhadap butir soal yang digunakan. Padahal, analisis butir soal merupakan langkah penting untuk menjamin bahwa setiap soal memiliki kemampuan mengukur capaian belajar secara tepat dan adil.

Kondisi ini menunjukkan adanya kesenjangan riset (research gap), terutama dalam konteks analisis butir soal pada mata pelajaran bahasa Arab di tingkat madrasah. Berbagai penelitian sebelumnya telah menelaah analisis butir soal pada mata pelajaran lain seperti matematika, IPA, maupun bahasa Indonesia, namun kajian serupa dalam konteks bahasa Arab masih relatif terbatas. Padahal, pembelajaran bahasa Arab memiliki karakteristik tersendiri, terutama dalam aspek morfologi dan sintaksis, yang menuntut instrumen evaluasi yang lebih akurat dan representatif terhadap kompetensi berbahasa peserta didik. Kurangnya penelitian di bidang ini berimplikasi pada minimnya referensi dan praktik analisis soal yang berbasis data empiris di lingkungan madrasah.

Oleh karena itu, penelitian ini hadir dengan kebaruan (novelty) pada fokus analisisnya terhadap kualitas butir soal pilihan ganda bahasa Arab di MTs al-Hidayah Batu, mencakup analisis tingkat kesukaran dan daya pembeda setiap butir soal. Melalui pendekatan kuantitatif, penelitian ini memberikan gambaran objektif mengenai sejauh mana soal-soal yang digunakan mampu membedakan siswa berkemampuan tinggi dan rendah, sekaligus mengukur apakah tingkat kesukarannya seimbang sesuai dengan desain pembelajaran. Hasil penelitian ini diharapkan tidak hanya memperkaya khazanah penelitian tentang evaluasi pembelajaran bahasa Arab, tetapi juga menjadi rujukan praktis bagi guru dalam menyusun dan memperbaiki soal evaluasi agar sesuai dengan prinsip evaluasi yang valid, reliabel, dan efektif.

Dengan demikian, penelitian ini berangkat dari kebutuhan mendesak untuk memperbaiki kualitas evaluasi pembelajaran di madrasah melalui analisis mendalam terhadap butir soal pilihan ganda. Fokus penelitian pada analisis empiris instrumen evaluasi bahasa Arab menjadi pembeda utama dari penelitian sebelumnya, sekaligus memberikan kontribusi nyata bagi peningkatan mutu penilaian hasil belajar di lembaga pendidikan Islam.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode analisis deskriptif. Pendekatan kuantitatif dipilih karena penelitian ini bertujuan untuk menganalisis data numerik yang berkaitan dengan tingkat kesukaran dan daya pembeda butir soal pilihan ganda pada mata pelajaran Bahasa Arab. Menurut (Nadirah et al., 2022), penelitian kuantitatif berlandaskan pada filsafat positivisme, menekankan pada pengujian teori melalui pengukuran variabel dengan data numerik, serta menggunakan analisis statistik untuk menilai hubungan antarvariabel atau kualitas instrumen penelitian. Dengan demikian, metode ini dianggap paling tepat untuk menilai validitas empiris dari butir soal yang diuji.

1. Desain Penelitian

Desain penelitian ini bersifat analisis butir soal (item analysis) yang berfokus pada dua indikator utama, yaitu tingkat kesukaran (difficulty index) dan daya pembeda (discrimination index). Analisis dilakukan untuk menilai sejauh mana butir soal dapat mengukur kemampuan peserta didik secara proporsional dan mampu membedakan antara siswa yang berkemampuan tinggi dan rendah. Pendekatan ini bertujuan menghasilkan gambaran objektif mengenai kualitas butir soal dalam konteks pembelajaran Bahasa Arab.

2. Populasi dan Sampel Penelitian

Populasi dalam penelitian ini adalah seluruh peserta didik kelas VIII MTs Al-Hidayah Kota Batu pada tahun ajaran 2024/2025. Penentuan populasi didasarkan pada pertimbangan bahwa siswa kelas VIII telah menempuh sebagian besar materi Bahasa Arab tingkat menengah dan telah mengikuti evaluasi sumatif yang memuat butir soal dalam bentuk pilihan ganda. Seluruh siswa kelas VIII dijadikan sampel penelitian (total sampling) karena jumlahnya relatif terjangkau dan sesuai untuk analisis butir soal yang membutuhkan representasi seluruh kelompok kemampuan siswa.

3. Instrumen Penelitian

Instrumen utama dalam penelitian ini adalah tes hasil belajar Bahasa Arab dalam bentuk soal pilihan ganda sebanyak 25 butir. Soal tersebut disusun berdasarkan silabus, Rencana Pelaksanaan Pembelajaran (RPP), dan buku ajar Bahasa Arab kelas VIII yang digunakan di madrasah. Setiap butir soal terdiri atas satu stem (pokok soal) dan empat alternatif jawaban, dengan satu jawaban yang benar dan tiga pengecoh (distraktor). Penyusunan soal mengacu pada indikator pencapaian kompetensi dari Kompetensi Inti (KI) dan Kompetensi Dasar (KD) yang berlaku dalam kurikulum madrasah. Sebelum digunakan dalam penelitian, soal telah melalui proses telaah guru mata pelajaran dan validasi isi untuk memastikan kesesuaian materi dan tingkat kesulitan yang proporsional.

4. Prosedur Pengumpulan Data

Pengumpulan data dilakukan melalui pelaksanaan tes tertulis kepada seluruh responden pada waktu yang telah dijadwalkan oleh pihak madrasah. Setiap peserta didik mengerjakan 25 butir soal dalam waktu yang sama dan kondisi pengawasan seragam agar hasil tes merefleksikan kemampuan sebenarnya. Setelah pelaksanaan tes, lembar jawaban dikoreksi dan diberi skor berdasarkan jumlah jawaban benar. Skor yang diperoleh tiap peserta didik kemudian diolah menjadi data numerik yang siap dianalisis untuk menentukan tingkat kesukaran dan daya pembeda masing-masing butir soal.

5. Teknik Analisis Data

Analisis data dilakukan secara kuantitatif menggunakan rumus statistik untuk menghitung dua aspek utama, yaitu tingkat kesukaran (P) dan daya pembeda (D) dari setiap butir soal.

a. Tingkat Kesukaran (Difficulty Index / P)

Tingkat kesukaran butir soal menunjukkan proporsi siswa yang mampu menjawab benar suatu soal. Nilai P dihitung menggunakan rumus:

$$P = \frac{B}{N}$$

di mana:
B= jumlah peserta yang menjawab benar,
N= jumlah seluruh peserta tes.

Interpretasi nilai P adalah sebagai berikut:

- 0.00 – 0.30 : Soal sukar
- 0.31 – 0.70 : Soal sedang
- 0.71 – 1.00 : Soal mudah

Butir soal yang baik biasanya memiliki tingkat kesukaran sedang, karena dapat menilai kemampuan siswa secara lebih representatif.

b. Daya Pembeda (Discrimination Index / D)

Daya pembeda menunjukkan kemampuan butir soal untuk membedakan antara siswa berkemampuan tinggi dan rendah. Nilai D dihitung dengan rumus:

$$D = P_A - P_B$$

di mana:
P_A= proporsi siswa kelompok atas yang menjawab benar,
P_B= proporsi siswa kelompok bawah yang menjawab benar.

Kriteria interpretasi nilai D adalah:

a. $\bar{D}$ 0.20 : Daya pembeda rendah
b. 0.21 – 0.40 : Daya pembeda sedang
c. 0.41 – 0.70 : Daya pembeda tinggi
d. $\bar{D}$ 0.70 : Daya pembeda sangat tinggi

Nilai D yang positif menunjukkan bahwa soal tersebut dapat membedakan kemampuan siswa dengan baik, sedangkan nilai D negatif menunjukkan bahwa soal tersebut tidak efektif karena lebih mudah dijawab oleh siswa berkemampuan rendah.

6. Validitas dan Reliabilitas

Untuk menjamin keakuratan hasil analisis, dilakukan uji validitas isi (content validity) dengan melibatkan guru ahli Bahasa Arab untuk menilai kesesuaian soal terhadap indikator kompetensi. Selain itu, dilakukan uji reliabilitas menggunakan rumus Kuder-Richardson 20 (KR-20) untuk menentukan konsistensi internal instrumen. Nilai reliabilitas yang tinggi menunjukkan bahwa butir soal memiliki kestabilan dan konsistensi hasil yang baik.

7. Langkah Analisis Lanjutan

Hasil analisis tingkat kesukaran dan daya pembeda disajikan dalam tabel distribusi yang menggambarkan kategori tiap butir soal. Soal-soal yang terlalu mudah, terlalu sukar, atau memiliki daya pembeda rendah kemudian diidentifikasi untuk direvisi. Hasil akhir dianalisis secara deskriptif guna memberikan rekomendasi terhadap kualitas butir soal dan efektivitasnya sebagai alat evaluasi pembelajaran Bahasa Arab.

HASIL DAN PEMBAHASAN

Analisis Tingkat Kesukaran Butir Soal

Analisis tingkat kesukaran butir soal merupakan langkah penting dalam mengevaluasi kualitas instrumen tes. Tujuan analisis ini adalah untuk mengetahui sejauh mana butir soal dapat membedakan peserta didik berdasarkan kemampuannya. Tingkat kesukaran tidak ditentukan oleh pandangan subjektif guru, melainkan oleh proporsi peserta didik yang mampu menjawab benar. Dengan demikian, semakin banyak peserta didik yang menjawab benar, semakin mudah soal tersebut, dan sebaliknya.

Tingkat kesukaran diukur menggunakan indeks kesukaran (P) yang memiliki rentang antara 0,00 sampai 1,00. Nilai P yang mendekati 0,00 menunjukkan soal tergolong sukar karena hampir tidak ada siswa yang menjawab benar, sedangkan nilai mendekati 1,00 menunjukkan soal tergolong mudah karena hampir seluruh siswa menjawab benar. Hasil perhitungan tingkat kesukaran dari 25 butir soal dapat dilihat pada tabel berikut:

Tabel 1. Distribusi Tingkat Kesukaran Butir Soal

Kategori	Rentang Nilai P	Jumlah Soal	Persentase	Nomor Soal
Sukar	$P \leq 0{,}30$	1	4%	8
Sedang	$0{,}30 < P \leq 0{,}70$	8	32%	2, 3, 5, 7, 10, 11, 16, 22
Mudah	$0{,}70 < P \leq 1{,}00$	16	64%	1, 4, 6, 9, 12, 13, 14, 15, 17, 18, 19, 20, 21, 23, 24, 25
Total		**25**	**100%**	

Sumber: Data diolah 2025

Berdasarkan hasil analisis pada Tabel 1, dari total 25 butir soal ujian Bahasa Arab, diperoleh 1 butir soal (4%) yang termasuk kategori sukar, 8 butir soal (32%) kategori sedang, dan 16 butir soal (64%) kategori mudah. Proporsi ini menunjukkan bahwa sebagian besar butir soal masih cenderung mudah bagi peserta didik. Hal ini perlu diperhatikan karena tes yang terlalu mudah tidak dapat menggambarkan kemampuan siswa secara menyeluruh.

Untuk melihat sejauh mana kesesuaian antara desain awal dan hasil aktual, dilakukan perbandingan antara proyeksi tingkat kesukaran yang direncanakan dengan realisasi di lapangan.

Tabel 2. Perbandingan Tingkat Kesukaran Proyeksi vs Realisasi

Kategori	Proyeksi Awal	Realisasi	Selisih
Sukar	2-3 soal (10%)	1 soal (4%)	-1 hingga -2 soal
Sedang	10 soal (40%)	8 soal (32%)	-2 soal
Mudah	12-13 soal (50%)	16 soal (64%)	+3 hingga +4 soal

Sumber: Data diolah 2025

Dari Tabel 2 dapat dilihat bahwa hasil realisasi tidak sepenuhnya sesuai dengan proyeksi awal. Jumlah soal kategori sukar dan sedang lebih sedikit dari yang diharapkan, sedangkan soal mudah justru lebih banyak. Secara khusus, terdapat kekurangan sekitar 2 soal dalam kategori sedang dan kelebihan sekitar 3–4 soal dalam kategori mudah. Ketidakseimbangan ini menunjukkan perlunya peninjauan ulang terhadap komposisi tingkat

kesukaran untuk mencapai distribusi ideal sesuai dengan standar pengujian yang proporsional.

Selanjutnya, untuk mengidentifikasi butir mana saja yang tidak sesuai dengan proyeksi awal, dilakukan analisis mendalam terhadap soal yang menunjukkan perbedaan signifikan antara tingkat kesukaran yang direncanakan dan hasil perhitungannya.

Tabel 3. Soal yang Tidak Sesuai Proyeksi

No. Soal	Proyeksi Awal	Hasil Analisis	Rekomendasi
8	Sukar	Sedang (P=0,27)	Tingkatkan kesulitan
12	Sedang	Mudah (P=0,73)	Tingkatkan kesulitan
13	Sedang	Mudah (P=0,93)	Tingkatkan kesulitan
14	Sedang	Mudah (P=0,80)	Tingkatkan kesulitan
22	Mudah	Sedang (P=0,40)	Turunkan kesulitan

Sumber: Data diolah 2025

Berdasarkan Tabel 3, terdapat lima butir soal yang tidak sesuai dengan rancangan awal tingkat kesukaran. Empat butir soal (nomor 8, 12, 13, dan 14) memiliki tingkat kesukaran lebih rendah dari yang diharapkan sehingga perlu dilakukan revisi untuk meningkatkan tingkat kesulitannya. Sementara itu, satu butir soal (nomor 22) justru lebih sulit dari yang diproyeksikan, sehingga perlu dilakukan penyederhanaan agar sesuai dengan kategori mudah.

Secara keseluruhan, hasil analisis tingkat kesukaran menunjukkan bahwa sebagian besar butir soal masih berada pada kategori mudah dan hanya sebagian kecil yang tergolong sedang dan sukar. Kondisi ini menandakan perlunya penyesuaian dalam perancangan soal agar distribusi tingkat kesukaran lebih proporsional dan dapat mengukur kemampuan siswa secara lebih akurat.

Analisis Daya Pembeda Butir Soal

Daya pembeda adalah kemampuan butir soal untuk membedakan antara peserta didik berkemampuan tinggi dengan yang berkemampuan rendah (Akhmadi, 2021). Daya pembeda dihitung berdasarkan perbedaan proporsi jawaban benar antara kelompok atas (siswa pandai) dan kelompok bawah (siswa kurang pandai). Nilai daya pembeda (D) yang positif menunjukkan bahwa soal dapat membedakan dengan baik, di mana siswa kelompok atas lebih banyak menjawab benar dibanding kelompok bawah. Sebaliknya, nilai D negatif menunjukkan daya pembeda yang terbalik (Qadir, Huda, & Hermina, 2024).

Untuk mengetahui sejauh mana kemampuan soal dalam membedakan peserta didik berdasarkan kemampuan akademiknya, dilakukan analisis distribusi daya pembeda terhadap 25 butir soal yang diujikan. Hasil perhitungan tersebut disajikan pada tabel berikut:

Tabel 4. Distribusi Daya Pembeda Butir Soal

Kategori	Rentang Nilai D	Jumlah Soal	Persentase	Nomor Soal
Sangat Rendah	$D \leq 0$	6	24%	1, 7, 10, 16, 22, 25
Rendah	$0 < D \leq 0{,}20$	1	4%	3
Sedang	$0{,}20 < D \leq 0{,}40$	4	16%	2, 5, 11, 23
Tinggi	$0{,}40 < D \leq 0{,}70$	9	36%	4, 6, 12, 14, 17, 19, 20
Sangat Tinggi	$0{,}70 < D \leq 1{,}00$	5	20%	9, 13, 15, 18, 21
Total		**25**	**100%**	

Sumber: Data diolah 2025

Berdasarkan Tabel 4 di atas, terlihat bahwa sebagian besar butir soal memiliki daya pembeda dalam kategori tinggi dan sangat tinggi. Sebanyak 9 soal (36%) termasuk kategori tinggi dan 5 soal (20%) termasuk kategori sangat tinggi, sehingga dapat disimpulkan bahwa 14 butir soal (56%) berfungsi dengan baik dalam membedakan siswa berkemampuan tinggi dan rendah. Sementara itu, terdapat 4 soal (16%) dengan kategori sedang, 1 soal (4%) berkategori rendah, dan 6 soal (24%) berkategori sangat rendah. Kondisi ini menunjukkan bahwa meskipun sebagian besar butir soal telah efektif, masih terdapat beberapa soal yang memerlukan perbaikan agar berfungsi optimal sebagai alat evaluasi pembelajaran.

Selanjutnya, untuk memperjelas butir-butir soal yang memiliki daya pembeda sangat rendah bahkan negatif (terbalik), dilakukan identifikasi lebih lanjut terhadap soal dengan nilai D negatif. Soal-soal tersebut dianggap tidak valid karena gagal membedakan siswa pandai dan kurang pandai. Hasilnya dapat dilihat pada tabel berikut:

Tabel 5. Soal dengan Indeks Diskriminasi Negatif

No. Soal	Nilai D	Tingkat Kesukaran	Interpretasi
1	-0,07	Mudah (0,73)	Lebih mudah bagi kelompok bawah
7	-0,07	Sedang (0,33)	Lebih mudah bagi kelompok bawah
10	-0,65	Sedang (0,33)	Sangat tidak efektif
16	-0,03	Sedang (0,67)	Lebih mudah bagi kelompok bawah
22	-0,03	Sedang (0,40)	Lebih mudah bagi kelompok bawah
25	-0,15	Mudah (0,93)	Lebih mudah bagi kelompok bawah

Sumber: Data diolah 2025

Tabel 5 menunjukkan bahwa terdapat enam butir soal dengan nilai diskriminasi negatif. Soal-soal tersebut cenderung lebih mudah dijawab benar oleh kelompok siswa berkemampuan rendah dibandingkan kelompok berkemampuan tinggi. Fenomena ini menandakan adanya ketidaksesuaian dalam konstruksi atau redaksi soal, yang dapat disebabkan oleh faktor ketidaktepatan dalam penyusunan indikator, konteks soal yang ambigu, atau ketidaksesuaian tingkat kesulitan dengan kemampuan siswa. Oleh karena itu, keenam butir soal ini sebaiknya direvisi atau diganti agar dapat berfungsi sebagai alat ukur yang valid dan reliabel dalam mengevaluasi capaian belajar siswa.

Hubungan Tingkat Kesukaran dan Daya Pembeda

Analisis butir soal dilakukan untuk mengetahui sejauh mana setiap butir dapat membedakan siswa berkemampuan tinggi dan rendah serta tingkat kesukarannya. Hasil analisis tingkat kesukaran (P) dan daya pembeda (D) disajikan pada Tabel 6 berikut:

Tabel 6. Rekapitulasi Analisis Butir Soal

No	Indeks Kesukaran (P)	Kategori Kesukaran	Indeks Diskriminasi (D)	Kategori Daya Pembeda	Status
1	0,73	Mudah	-0,07	Sangat Rendah	Revisi
2	0,53	Sedang	0,34	Sedang	Baik
3	0,67	Sedang	0,01	Rendah	Revisi
4	0,80	Mudah	0,59	Tinggi	Baik
5	0,60	Sedang	0,36	Sedang	Baik
6	0,87	Mudah	0,63	Tinggi	Baik
7	0,33	Sedang	-0,07	Sangat Rendah	Revisi
8	0,27	Sukar	0,29	Sedang	Perbaiki
9	0,93	Mudah	0,79	Sangat Tinggi	Sangat Baik

10	0,33	Sedang	-0,65	Sangat Rendah	Ganti
11	0,67	Sedang	0,32	Sedang	Baik
12	0,73	Mudah	0,54	Tinggi	Baik
13	0,93	Mudah	0,79	Sangat Tinggi	Sangat Baik
14	0,80	Mudah	0,44	Tinggi	Baik
15	0,93	Mudah	0,79	Sangat Tinggi	Sangat Baik
16	0,67	Sedang	-0,03	Sangat Rendah	Revisi
17	0,80	Mudah	0,44	Tinggi	Baik
18	0,93	Mudah	0,79	Sangat Tinggi	Sangat Baik
19	0,80	Mudah	0,44	Tinggi	Baik
20	0,73	Mudah	0,68	Tinggi	Baik
21	0,87	Mudah	0,75	Sangat Tinggi	Sangat Baik
22	0,40	Sedang	-0,03	Sangat Rendah	Revisi
23	0,73	Mudah	0,40	Tinggi	Baik
24	0,73	Mudah	0,00	Sangat Rendah	Revisi
25	0,93	Mudah	-0,15	Sangat Rendah	Revisi

Sumber: Data diolah 2025

Dari Tabel 6 terlihat bahwa terdapat variasi yang cukup signifikan antara indeks kesukaran dan daya pembeda setiap butir soal. Sebagian besar soal berada pada kategori mudah hingga sedang dengan nilai indeks kesukaran (P) antara 0,53 sampai 0,93. Namun, terdapat pula beberapa butir soal dengan nilai daya pembeda negatif, yang menunjukkan bahwa butir tersebut tidak mampu membedakan siswa kelompok atas dan bawah secara efektif.

Hasil analisis menunjukkan adanya hubungan yang konsisten antara tingkat kesukaran dan daya pembeda butir soal, yakni:

1. Soal dengan P mendekati 0 atau 1 cenderung memiliki daya pembeda rendah (D mendekati 0). Contoh: soal nomor 24 (P=0,73; D=0,00) dan soal nomor 25 (P=0,93; D=-0,15).

2. Soal dengan tingkat kesukaran sedang (P sekitar 0,40-0,60) umumnya memiliki potensi daya pembeda lebih baik. Contoh: soal nomor 2 (P=0,53; D=0,34) dan soal nomor 5 (P=0,60; D=0,36).

3. Soal terlalu mudah dapat menghasilkan daya pembeda negatif karena hampir semua siswa (termasuk kelompok bawah) dapat menjawab dengan benar.

Temuan ini memperkuat prinsip bahwa butir soal yang ideal adalah yang memiliki tingkat kesukaran sedang dan daya pembeda tinggi, karena dapat menyeleksi kemampuan siswa secara proporsional. Berdasarkan hasil tersebut, dilakukan pengelompokan tindak lanjut terhadap setiap kategori soal untuk menentukan langkah perbaikan atau pemanfaatan selanjutnya. Rincian rekomendasi tindak lanjut tersebut ditampilkan pada Tabel 7 berikut.

Tabel 7. Rekomendasi Tindak Lanjut

Kategori Soal	Jumlah	Rekomendasi
Sangat Baik (Kesukaran sedang-mudah, Daya pembeda tinggi-sangat tinggi)	5	Pertahankan dan masukkan ke bank soal
Baik (Daya pembeda sedang-tinggi)	11	Pertahankan dengan perbaikan minor
Perlu Perbaikan (Daya pembeda rendah atau kesukaran tidak sesuai)	2	Revisi substansi soal

Perlu Revisi Total (Daya pembeda negatif atau sangat rendah)	7	Revisi total atau ganti dengan soal baru

Sumber: Data diolah 2025

Tabel 7 menunjukkan bahwa dari keseluruhan butir soal, terdapat 5 soal yang tergolong *sangat baik* dan dapat langsung dimasukkan ke bank soal, 11 soal tergolong *baik* dengan perbaikan minor, serta 2 soal yang perlu direvisi substansinya karena memiliki daya pembeda rendah. Sementara itu, 7 butir soal lainnya dikategorikan perlu *revisi total* atau bahkan diganti dengan soal baru karena menunjukkan daya pembeda negatif atau sangat rendah.

Dengan demikian, hasil analisis ini memberikan dasar yang kuat untuk penyusunan ulang instrumen evaluasi agar lebih valid dan reliabel dalam mengukur kemampuan peserta didik secara proporsional.

Pembahasan

Hasil penelitian ini menunjukkan bahwa tingkat kesukaran dan daya pembeda butir soal Bahasa Arab di MTs Al-Hidayah Kota Batu memiliki variasi yang cukup signifikan dan belum sepenuhnya memenuhi proporsi ideal sebagaimana standar evaluasi pendidikan. Berdasarkan analisis pada Tabel 1, diketahui bahwa sebagian besar butir soal (64%) termasuk kategori mudah, sedangkan hanya 32% berada pada kategori sedang dan 4% tergolong sukar. Dominasi butir soal kategori mudah ini menunjukkan bahwa sebagian besar peserta didik dapat menjawab dengan benar, sehingga kemampuan kognitif tingkat tinggi belum sepenuhnya terukur.

Temuan ini sejalan dengan pendapat (Kaka et al., 2024) bahwa soal yang terlalu mudah atau terlalu sukar tidak dapat berfungsi optimal dalam mengukur perbedaan kemampuan peserta didik. Soal yang baik umumnya memiliki indeks kesukaran sedang karena mampu memberikan variasi hasil antara siswa berkemampuan tinggi dan rendah. Hasil penelitian ini juga mendukung temuan (Fatimah & Alfath, 2019) yang menyatakan bahwa komposisi ideal tingkat kesukaran dalam tes sebaiknya mencakup 30% soal mudah, 50% sedang, dan 20% sukar agar distribusi hasil lebih representatif terhadap seluruh peserta.

Perbandingan antara desain awal dan hasil realisasi pada Tabel 2 menunjukkan bahwa jumlah soal sedang dan sukar lebih sedikit dari yang direncanakan, sementara soal mudah justru meningkat. Hal ini menandakan adanya ketidaktepatan dalam proses penyusunan atau pengujian awal soal. Sejalan dengan (Susanto et al., 2015), perbedaan antara proyeksi dan realisasi tingkat kesukaran sering kali disebabkan oleh kurangnya uji coba awal (try out) dan analisis empiris terhadap butir soal sebelum digunakan dalam ujian sesungguhnya. Oleh karena itu, perencanaan tingkat kesukaran perlu disertai dengan verifikasi empiris agar soal yang digunakan sesuai dengan karakteristik peserta didik.

Berdasarkan analisis lanjut pada Tabel 3, ditemukan lima butir soal yang tidak sesuai dengan rancangan awal, yakni nomor 8, 12, 13, 14, dan 22. Sebagian besar butir tersebut memerlukan peningkatan atau penurunan tingkat kesulitan agar sesuai dengan kategori yang diharapkan. Fenomena ini memperkuat pandangan (Akhmadi, 2021) bahwa perbedaan hasil antara rancangan teoritis dan empiris merupakan indikasi penting bagi guru untuk melakukan peninjauan ulang terhadap indikator soal dan redaksi kalimat yang digunakan.

Selanjutnya, hasil analisis daya pembeda pada Tabel 4 menunjukkan bahwa 56% dari keseluruhan butir soal tergolong baik (kategori tinggi dan sangat tinggi), sedangkan sisanya masih menunjukkan daya pembeda sedang hingga sangat rendah. Nilai daya pembeda yang tinggi menandakan bahwa soal tersebut efektif dalam membedakan peserta didik berkemampuan tinggi dan rendah, sesuai dengan teori (Qadir et al., 2024) yang menekankan bahwa nilai diskriminasi positif mengindikasikan validitas instrumen dalam mengukur perbedaan kemampuan kognitif. Namun, terdapat enam butir soal dengan nilai daya pembeda negatif sebagaimana ditunjukkan pada Tabel 5.

Soal-soal dengan nilai diskriminasi negatif menunjukkan bahwa siswa berkemampuan rendah justru lebih banyak menjawab benar dibandingkan siswa berkemampuan tinggi. Kondisi ini dapat disebabkan oleh ketidaksesuaian konteks soal, kesalahan kunci jawaban, atau redaksi soal yang menimbulkan multitafsir. Menurut (Fatimah & Alfath, 2019), daya pembeda negatif mengindikasikan adanya butir soal yang perlu direvisi secara mendasar karena tidak mampu menjalankan fungsi evaluatif secara proporsional. Hasil ini juga senada dengan penelitian (Basri et al., 2025) yang menemukan bahwa soal dengan redaksi ambigu sering kali menghasilkan korelasi negatif antara kemampuan siswa dan hasil jawaban.

Hubungan antara tingkat kesukaran dan daya pembeda sebagaimana ditampilkan pada Tabel 6 memperlihatkan pola yang konsisten dengan teori pengukuran klasik. Soal yang terlalu mudah (nilai P mendekati 1,00) atau terlalu sukar (nilai P mendekati 0,00) cenderung memiliki daya pembeda rendah atau bahkan negatif. Sementara itu, soal dengan tingkat kesukaran sedang (P sekitar 0,40–0,60) menunjukkan daya pembeda yang lebih optimal. Pola ini sesuai dengan pendapat (Ainin, 2006) yang menyatakan bahwa butir soal dengan tingkat kesukaran sedang memiliki potensi terbesar dalam memberikan informasi diferensiasi kemampuan antar peserta tes.

Untuk menindaklanjuti hasil tersebut, dilakukan pengelompokan rekomendasi sebagaimana tercantum dalam Tabel 7. Dari total 25 butir soal, terdapat 5 butir yang dikategorikan sangat baik dan dapat langsung dimasukkan ke bank soal, 11 butir tergolong baik dengan perbaikan minor, 2 butir perlu revisi substansi, dan 7 butir perlu revisi total. Distribusi ini menandakan bahwa lebih dari separuh butir soal telah memenuhi kriteria kualitas yang baik, namun masih ada sepertiga yang perlu ditingkatkan validitasnya.

Temuan ini sejalan dengan hasil penelitian (Maulana & Sanusi, 2020) yang menegaskan pentingnya uji empiris terhadap setiap butir soal untuk memastikan reliabilitas dan validitasnya dalam konteks pembelajaran bahasa Arab. Selain itu, penelitian ini memperluas hasil studi (Somahhida et al., 2022) yang menunjukkan bahwa efektivitas evaluasi pembelajaran bahasa bergantung pada keseimbangan antara tingkat kesukaran dan daya pembeda butir soal.

Dengan demikian, secara teoretis hasil penelitian ini mendukung teori pengukuran klasik (Classical Test Theory/CTT) yang menyatakan bahwa kualitas tes ditentukan oleh keseimbangan antara tingkat kesukaran dan daya pembeda. Secara empiris, penelitian ini memperkaya kajian terdahulu dengan bukti nyata bahwa pada mata pelajaran Bahasa Arab, dominasi soal mudah berimplikasi pada menurunnya daya seleksi kemampuan siswa. Oleh karena itu, disarankan agar guru bahasa Arab melakukan analisis butir soal secara berkala setelah setiap ujian, agar instrumen evaluasi yang digunakan selalu valid, reliabel, dan sejalan dengan prinsip pengukuran pendidikan modern.

KESIMPULAN

Berdasarkan hasil analisis terhadap butir soal pilihan ganda mata pelajaran Bahasa Arab di MTs Al-Hidayah Kota Batu, dapat disimpulkan bahwa kualitas instrumen tes secara umum telah memenuhi sebagian kriteria evaluasi yang baik, namun masih memerlukan perbaikan pada beberapa aspek. Analisis tingkat kesukaran menunjukkan bahwa sebagian besar butir soal berada pada kategori mudah (64%), sedangkan hanya 32% berkategori sedang dan 4% tergolong sukar. Kondisi ini menandakan bahwa komposisi soal belum seimbang dan cenderung terlalu mudah bagi peserta didik, sehingga kurang optimal dalam menyeleksi kemampuan secara menyeluruh.

Sementara itu, hasil analisis daya pembeda menunjukkan bahwa 56% butir soal telah berfungsi dengan baik dalam membedakan kemampuan siswa, sedangkan sisanya masih memiliki daya pembeda rendah hingga negatif. Terdapat enam butir soal dengan nilai diskriminasi negatif, yang mengindikasikan bahwa butir tersebut tidak efektif dan perlu direvisi total. Hasil ini memperlihatkan bahwa sebagian butir soal belum mampu berfungsi secara optimal sebagai alat ukur yang valid dan reliabel.

Hubungan antara tingkat kesukaran dan daya pembeda menunjukkan pola yang konsisten dengan teori pengukuran klasik, yakni butir soal dengan tingkat kesukaran sedang memiliki daya pembeda yang lebih baik dibandingkan soal yang terlalu mudah atau terlalu sukar. Temuan ini menegaskan pentingnya keseimbangan komposisi soal dalam penyusunan instrumen evaluasi.

Secara keseluruhan, dari 25 butir soal yang dianalisis, terdapat 5 soal dengan kategori sangat baik yang dapat dimasukkan ke bank soal, 11 soal berkategori baik dengan perbaikan minor, 2 soal perlu revisi substansi, dan 7 soal perlu revisi total. Oleh karena itu, guru diharapkan melakukan analisis butir soal secara berkala dan berbasis data empiris agar kualitas evaluasi pembelajaran semakin meningkat.

Penelitian ini menegaskan bahwa analisis empiris terhadap butir soal tidak hanya berfungsi sebagai evaluasi teknis, tetapi juga sebagai strategi pedagogis untuk meningkatkan validitas, reliabilitas, dan efektivitas proses pembelajaran Bahasa Arab di madrasah.

DAFTAR REFERENSI

Ainin, M. dk. (2006). *Evaluasi dalam pembelajaran bahasa Arab* (Cet. I). Misykat.

Akhmadi, M. N. (2021). Analisis butir soal evaluasi tema 1 kelas 4 sdn plumbungan menggunakan program anates. *Ed-Humanistics: Jurnal Ilmu Pendidikan*, *6*(1), 799–806.

Basri, M. B., Sultan, S., Rapi, M., Baharman, B., & Sakaria, S. (2025). Meningkatkan Kompetensi Evaluasi Pembelajaran melalui Pelatihan Analisis Butir Soal bagi Mahasiswa Calon Guru. *MALLOMO: Journal of Community Service*, *5*(2), 633–639.

Faishal. (2017). Integrasi Ilmu Dalam Pendidikan. *Ta'dibi : Jurnal Prodi Manajemen Pendidikan Islam*, *VI*(2), 104–123.

Fatimah, L. U., & Alfath, K. (2019). Analisis kesukaran soal, daya pembeda dan fungsi distraktor. *AL-MANAR: Jurnal Komunikasi Dan Pendidikan Islam*, *8*(2), 37–64.

Hidayat, T., Rizal, A. S., & Fahrudin, F. (2018). Pendidikan Dalam Perspektif Islam Dan

Peranannya Dalam Membina Kepribadian Islami. *Jurnal MUDARRISUNA: Media Kajian Pendidikan Agama Islam*, *8*(2), 211–219.

Kaka, L., Bano, V. O., & Njoeroemana, Y. (2024). Efektivitas Analisis Butir Soal Pilihan Ganda Menggunakan Aplikasi Anates di SMPN 2 Kanatang. *Jurnal Inovasi Penelitian*, *4*(9), 1441–1450.

Mar, N. A., & Hilmi, D. (2021). Manajemen program pembelajaran bahasa Arab pada anak prasekolah Yayasan PAUD Sultan Qaimuddin di Kendari. *Jurnal Akuntabilitas Manajemen Pendidikan*, *9*(1), 1–10.

Maulana, D., & Sanusi, A. (2020). Analisis Butir Soal Bahasa Arab Ujian Akhir Madrasah Bersama Daerah (UAMBD) Madrasah Ibtidaiyah Tahun 2017-2018. *Ta'lim Al-'Arabiyyah: Jurnal Pendidikan Bahasa Arab & Kebahasaaraban*, *4*(1), 12–24.

Nadirah, S. P., Pramana, A. D. R., & Zari, N. (2022). *metodologi penelitian kualitatif, kuantitatif, mix method (mengelola Penelitian Dengan Mendeley dan Nvivo)*. CV. Azka Pustaka.

Nadya, A., Devia, D., & Gusmaneli, G. (2024). Hakikat Evaluasi (Pengertian Pengukuran, Penilaian, Evaluasi; Fungsi & Tujuan Penilaian, Ciri-Ciri Penilaian Pendidikan). *Jurnal Manajemen Dan Pendidikan Agama Islam*, *2*(2), 228–233.

Oktaviani, K. S. (2017). Bentuk tes objektif dan kecemasan pada pembelajaran membaca huruf hiragana bahasa Jepang. *Jurnal Evaluasi Pendidikan*, *8*(1), 455810.

Qadir, A., Huda, N., & Hermina, D. (2024). Analisis Butir Tes: Tingkat Kesukaran, Daya Pembeda Dan Efektivitas Pengecoh. *Al-Furqan: Jurnal Agama, Sosial, Dan Budaya*, *3*(3), 1450–1467.

Ramadani, E. N., & Handayani, D. F. (2024). Instrumen Penilaian Hasil Pembelajaran Kognitif Pada Tes Objektif. *Jurnal Pendidikan Dan Ilmu Sosial (Jupendis)*, *2*(4), 86–96.

Somahhida, N. G., Makruf, I., & Al-Alwany, N. (2022). Multiple Choice Objective Test Arabic Subject: Analysis & Implementation of the Edmodo Application/Tes Objektif Pilihan Ganda Mata Pelajaran Bahasa Arab: Analisis & Implementasi pada Aplikasi Edmodo. *ATHLA: Journal of Arabic Teaching, Linguistic and Literature*, *3*(2), 162–176.

Susanto, H., Rinaldi, A., & Novalia, N. (2015). Analisis validitas reliabilitas tingkat kesukaran dan daya beda pada butir soal ujian akhir semester ganjil mata pelajaran Matematika kelas XII IPS di SMA Negeri 12 Bandar Lampung tahun ajaran 2014/2015. *Al-Jabar: Jurnal Pendidikan Matematika*, *6*(2), 203–218.