EL-IBTIKAR
Jurnal Pendidikan Bahasa Arab
Vol. 1 No. 1, June 2026, pp.113-121
https://journal1.uinssc.ac.id/index.php/ibtikar/article/view/23856
DOI: https://doi.org/10.24235/r8k4qq42

Item Analysis of Arabic Language Test Questions Based on Difficulty Level, Discrimination Index, and Distractor Effectiveness

Anisah Firdausi Nuzula[1*], Maftukh Ghulam Mursyidan[2], M.Baihaqi[3]
[1 2 3] Universitas Islam Negri Sunan Ampel Surabaya
* Penulis Korespondensi: anisahfirdausi07@gmail.com

ABSTRAC

This study aims to analyze the quality of Arabic language test items based on difficulty level, discrimination index, and distractor effectiveness in formative assessment. The research employed a descriptive quantitative approach involving 30 students of an intensive Arabic program. The instrument consisted of 20 multiple-choice items covering grammar (qawā'id) and text comprehension. Data were analyzed using Classical Test Theory with the assistance of Microsoft Excel and SPSS. The results showed that 14 items were valid while 6 items were invalid. The reliability coefficient (Cronbach's Alpha) was 0.716, indicating good internal consistency. The difficulty level analysis revealed that 80% of the items were categorized as easy and 20% as moderate, with no difficult items found. In terms of discrimination index, most items were classified as good to very good, indicating their effectiveness in distinguishing students' abilities. Distractor analysis showed that most distractors functioned adequately, although some items had ineffective distractors due to low selection rates. These findings indicate that although the test instrument is generally reliable and has good discrimination power, improvements are needed in balancing item difficulty and enhancing distractor quality. Therefore, the development of more varied and higher-order thinking (HOTS)-based items is recommended to improve the overall quality of Arabic language assessment.

Key words: *item analysis, Arabic language test, difficulty level, discrimination index, distractor effectiveness*

ABSTRACT

Penelitian ini bertujuan untuk menganalisis kualitas butir soal Bahasa Arab berdasarkan tingkat kesukaran, daya pembeda, dan efektivitas distraktor pada asesmen formatif. Penelitian menggunakan pendekatan kuantitatif deskriptif dengan melibatkan 30 mahasiswa program intensif Bahasa Arab. Instrumen penelitian berupa 20 butir soal pilihan ganda yang mencakup aspek qawā'id dan pemahaman teks. Analisis data dilakukan menggunakan teori tes klasik dengan bantuan Microsoft Excel dan SPSS. Hasil penelitian menunjukkan bahwa 14 butir soal dinyatakan valid dan 6 butir tidak valid. Nilai reliabilitas sebesar 0,716 menunjukkan konsistensi internal yang baik. Analisis tingkat kesukaran menunjukkan bahwa 80% soal termasuk kategori mudah dan 20% kategori sedang, tanpa ditemukan soal kategori sulit. Dari aspek daya pembeda, sebagian besar soal berada pada kategori baik hingga sangat baik. Analisis distraktor menunjukkan bahwa sebagian besar pengecoh telah berfungsi dengan baik, meskipun masih terdapat beberapa yang kurang efektif. Temuan ini menunjukkan bahwa instrumen sudah cukup reliabel dan memiliki daya pembeda yang baik, namun masih perlu perbaikan pada distribusi tingkat kesukaran dan kualitas distraktor. Oleh karena itu, diperlukan pengembangan soal yang lebih bervariasi dan berbasis HOTS untuk meningkatkan kualitas evaluasi pembelajaran Bahasa Arab.

Kata kunci: *analisis butir soal, Bahasa Arab, tingkat kesukaran, daya pembeda, distraktor*

ABSTRAK

تهدف هذه الدراسة إلى تحليل جودة بنود اختبار اللغة العربية بناءً على مستوى الصعوبة، ومعامل التمييز، وفعالية المشتتات في التقييم التكويني. استخدمت الدراسة المنهج الكمي الوصفي بمشاركة 30 طالبًا من برنامج اللغة العربية المكثف. تألفت أداة البحث من 20 سؤالًا من نوع الاختيار من متعدد تغطي قواعد اللغة (القواعد) وفهم النصوص. تم تحليل البيانات باستخدام نظرية الاختبار الكلاسيكية بمساعدة برنامج Microsoft Excel وSPSS. أظهرت النتائج أن 14 بندًا كانت صالحة و6 بنود غير صالحة. بلغ معامل الثبات (ألفا كرونباخ) 0.716 مما يدل على اتساق داخلي جيد. وأظهرت نتائج تحليل مستوى الصعوبة أن 80% من الأسئلة كانت سهلة و20% متوسطة، دون وجود أسئلة صعبة.

أما من حيث معامل التمييز، فقد كانت معظم البنود في فئة جيدة إلى جيدة جدًا، مما يدل على قدرتها على التمييز بين مستويات الطلاب. كما أظهر تحليل المشتتات أن معظمها يعمل بشكل جيد، رغم وجود بعض المشتتات غير الفعالة. وتشير هذه النتائج إلى أن الأداة موثوقة بشكل عام، إلا أنها تحتاج إلى تحسين في توزيع مستوى الصعوبة وجودة المشتتات. لذلك، يُوصى بتطوير أسئلة أكثر تنوعًا تعتمد على مهارات التفكير العليا (HOTS) لتحسين جودة التقييم في تعلم اللغة العربية.

الكلمات الرئيسية: تحليل البنود، اختبار اللغة العربية، مستوى الصعوبة، معامل التمييز، المشتتات

Received: April 2, 2026	*Revised: April 28, 2026*	*Accepted:May 10, 2026*	*Published:June 17, 2026*
Date	*date*	*date*	*Date*

Citation (APA Style): Nuzula, A.F, et, al. (2026). Item Analysis of Arabic Language Test Questions Based on Difficulty Level, Discrimination Index, and Distractor Effectiveness. oleh staf editorial selama produksi

PENDAHULUAN

Evaluasi pembelajaran Bahasa Arab di Indonesia, khususnya di lingkungan perguruan tinggi, pada umumnya telah menjadi bagian penting dalam proses pembelajaran, terutama dalam program intensif Bahasa Arab di fakultas keagamaan seperti Fakultas Tarbiyah dan Ilmu Keguruan. Evaluasi ini umumnya dilakukan melalui berbagai bentuk asesmen, baik formatif maupun sumatif, yang bertujuan untuk mengukur kemampuan mahasiswa dalam memahami aspek kebahasaan seperti kosakata, tata bahasa, dan pemahaman teks. Namun, dalam praktiknya, penyusunan instrumen evaluasi masih cenderung berfokus pada penyelesaian materi tanpa diiringi dengan analisis kualitas butir soal secara mendalam.(Said et al., 2025) Hal ini menyebabkan instrumen yang digunakan belum sepenuhnya mampu menggambarkan kemampuan mahasiswa secara akurat dan komprehensif.

Secara ideal, asesmen formatif seharusnya dirancang sebagai alat untuk memantau dan meningkatkan proses pembelajaran secara berkelanjutan.(Harsono et al., 2024) Soal yang berkualitas hendaknya memiliki tingkat kesukaran yang seimbang, mampu membedakan peserta didik berdasarkan tingkat kemampuannya, serta dilengkapi dengan distraktor yang berfungsi secara optimal. Untuk mencapai kondisi tersebut, diperlukan penerapan analisis butir soal secara kuantitatif dan kualitatif setelah pelaksanaan tes. Solusi yang dapat ditawarkan antara lain adalah penggunaan teknik analisis klasik (classical test theory), pelatihan penyusunan soal bagi dosen, serta pengembangan bank soal berbasis analisis empiris yang terstandar.(Asria, 2023)

Sejumlah penelitian dalam enam tahun terakhir telah menegaskan pentingnya analisis butir soal dalam meningkatkan kualitas asesmen. Penelitian oleh Yusron (Yusron et al., 2020) menunjukkan bahwa analisis tingkat kesukaran dan daya pembeda dapat meningkatkan validitas tes secara signifikan. Studi oleh Naja (Naja et al., 2025) menemukan Instrumen tes diagnostik berbentuk *four-tier multiple choice* yang dikembangkan terbukti memiliki tingkat validitas yang baik, reliabel, praktis, serta efektif dalam mengidentifikasi tingkat pemahaman konsep, miskonsepsi, dan kesalahan siswa. Penelitian oleh Keysa (Maria et al., 2025) menunjukkan bahwa penilaian ranah kognitif dalam pembelajaran Bahasa Arab masih didominasi oleh pengukuran kemampuan berpikir tingkat rendah. Selanjutnya, penelitian oleh Neti (Hartati et al., 2019) mengungkapkan bahwa penerapan analisis butir soal secara sistematis dapat meningkatkan kualitas evaluasi pembelajaran. Terakhir, penelitian oleh Mashall (Nadapdap et al., 2025) menegaskan bahwa asesmen formatif berbasis analisis empiris mampu memberikan umpan balik yang lebih akurat bagi pengajar dan mahasiswa.

Penelitian ini menawarkan inovasi dengan mengintegrasikan analisis butir soal secara komprehensif pada asesmen formatif Bahasa Arab, khususnya pada program intensif di lingkungan UIN Sunan Ampel Surabaya. Inovasi yang diusulkan tidak hanya berfokus pada perhitungan statistik tingkat kesukaran dan daya pembeda, tetapi juga pada evaluasi kualitas distraktor secara mendalam dalam konteks linguistik Bahasa Arab. Selain itu, penelitian ini berupaya mengembangkan model analisis yang dapat digunakan sebagai acuan dalam penyusunan soal pemahaman yang lebih berkualitas, relevan, dan sesuai dengan karakteristik mahasiswa.

Penelitian ini bertujuan untuk mengkaji kualitas butir soal pemahaman Bahasa Arab ditinjau dari tingkat kesukaran, daya pembeda, dan efektivitas distraktor pada asesmen formatif

di Fakultas Tarbiyah dan Ilmu Keguruan UIN Sunan Ampel Surabaya. Selain itu, penelitian ini bertujuan untuk mengidentifikasi kelemahan butir soal yang digunakan serta memberikan rekomendasi perbaikan guna meningkatkan kualitas asesmen. Dengan demikian, hasil penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan evaluasi pembelajaran Bahasa Arab yang lebih valid, reliabel, dan efektif.

METODE

Penelitian ini menggunakan pendekatan kuantitatif deskriptif untuk menganalisis kualitas butir soal Bahasa Arab ditinjau dari aspek tingkat kesukaran, daya pembeda, dan efektivitas distraktor.(Saputra et al., 2022) Penelitian dilakukan pada mahasiswa program intensif Bahasa Arab di Fakultas Tarbiyah dan Ilmu Keguruan UIN Sunan Ampel Surabaya yang mengikuti asesmen formatif. Subjek penelitian ditentukan secara purposive, yaitu mahasiswa yang mengikuti tes secara lengkap dan memiliki data respons yang valid. Instrumen penelitian berupa soal pilihan ganda yang mencakup aspek pemahaman kosakata, struktur bahasa, dan pemahaman teks. Instrumen tersebut ditelaah terlebih dahulu untuk memastikan kesesuaian dengan tujuan pembelajaran, kejelasan bahasa, dan kelayakan isi.(Alfinnas, 2026) Data penelitian diperoleh melalui dokumentasi hasil tes mahasiswa berupa lembar jawaban atau file digital, kemudian dilakukan pengkodean jawaban berdasarkan kunci yang telah ditetapkan.

Prosedur penelitian meliputi pengumpulan data, verifikasi kelengkapan respons, pengolahan data, serta analisis butir soal. Pengolahan data dilakukan menggunakan bantuan Microsoft Excel dan IBM SPSS Statistics untuk mempermudah perhitungan dan rekapitulasi data.(Radha, 2025) Analisis butir soal dilakukan berdasarkan pendekatan teori tes klasik dengan mengkaji tingkat kesukaran, daya pembeda, dan efektivitas distraktor pada setiap butir soal. Kriteria yang digunakan meliputi klasifikasi tingkat kesukaran (mudah, sedang, sulit), daya pembeda (rendah hingga sangat baik), serta distraktor yang dinilai efektif apabila dipilih oleh peserta dengan proporsi tertentu dan mampu mengecoh peserta berkemampuan rendah. Hasil analisis digunakan untuk menentukan kualitas setiap butir soal apakah layak digunakan, perlu direvisi, atau harus dibuang, serta dilakukan verifikasi ulang guna menjamin konsistensi dan keakuratan data

TEMUAN DAN DISKUSI

Temuan

Deskripsi data penelitian ini mencakup 30 mahasiswa pada program intensif Bahasa Arab di Fakultas Tarbiyah dan Ilmu Keguruan. Instrumen yang digunakan berupa 20 butir soal pilihan ganda dalam bentuk asesmen formatif yang mengukur pemahaman mahasiswa pada materi qawā‘id, khususnya konsep *mubtada–khabar* dan *na‘t–man‘ūt*. Data yang dianalisis berupa respons jawaban mahasiswa terhadap seluruh butir soal.

1. Hasil Uji Validitas Butir Soal
   Hasil uji validitas menggunakan korelasi Pearson Product Moment menunjukkan bahwa sebagian besar butir soal telah memenuhi kriteria validitas (r hitung > r tabel = 0,361). Rincian hasil uji validitas disajikan pada Tabel 1.

Tabel 1. Hasil Uji Validitas Butir Soal

No Soal	r hitung	Keterangan
1	0,527	Valid

2	0,527	Valid
3	0,479	Valid
4	0,064	Tidak Valid
5	0,642	Valid
6	0,323	Tidak Valid
7	0,186	Tidak Valid
8	0,482	Valid
9	0,479	Valid
10	-0,146	Tidak Valid
11	0,473	Valid
12	0,550	Valid
13	0,167	Tidak Valid
14	0,711	Valid
15	0,373	Valid
16	0,033	Tidak Valid
17	0,592	Valid
18	0,532	Valid
19	0,605	Valid
20	0,541	Valid

Berdasarkan tabel tersebut, sebanyak 14 dari 20 butir soal dinyatakan valid, sedangkan 6 butir lainnya tidak memenuhi kriteria. Temuan ini menunjukkan bahwa secara umum instrumen telah memiliki tingkat validitas yang memadai, meskipun masih diperlukan revisi pada beberapa butir yang tidak valid.

2. Hasil Uji Reliabilitas

   Hasil uji reliabilitas menunjukkan nilai Cronbach's Alpha sebesar 0,716. Nilai ini berada di atas batas minimal 0,70, sehingga instrumen dapat dikategorikan memiliki konsistensi internal yang baik. Dengan demikian, instrumen dinilai reliabel dan layak digunakan dalam pengukuran.

3. Tingkat Kesukaran Soal

   Distribusi tingkat kesukaran butir soal disajikan pada Tabel 2.

Tabel 2. Distribusi Tingkat Kesukaran Soal

Kategori	Jumlah Soal	Persentase
Mudah	16	80%
Sedang	4	20%
Sulit	0	0%

Hasil analisis menunjukkan bahwa mayoritas soal berada pada kategori mudah. Hal ini mengindikasikan bahwa tingkat kesulitan instrumen relatif rendah, sehingga kurang optimal dalam mengukur variasi kemampuan mahasiswa secara menyeluruh

4. Daya Pembeda Soal

   Hasil analisis daya pembeda disajikan pada Tabel 3.

Tabel 3. Distribusi Daya Pembeda Soal

Kategori	Jumlah Soal	Persentase
Sangat Baik	14	70%
Baik	5	25%
Cukup	1	5%
Kurang	0	0%

Secara umum, butir soal memiliki daya pembeda yang baik. Hal ini menunjukkan bahwa instrumen mampu membedakan mahasiswa berkemampuan tinggi dan rendah secara efektif.

5. Efektivitas Distraktor

   Hasil analisis menunjukkan bahwa sebagian besar distraktor telah berfungsi dengan baik, terutama pada butir soal dengan daya pembeda tinggi. Namun, pada soal yang tergolong mudah, ditemukan beberapa distraktor yang kurang efektif karena jarang dipilih oleh mahasiswa. Kondisi ini menunjukkan perlunya perbaikan dalam penyusunan opsi jawaban agar lebih representatif terhadap kemungkinan kesalahan mahasiswa

Diskusi

Hasil penelitian menunjukkan bahwa dari 20 butir soal, sebanyak 14 butir dinyatakan valid karena memiliki nilai r hitung yang melebihi r tabel (0,361), sehingga mampu merepresentasikan konstruk yang diukur, sedangkan 6 butir lainnya tidak valid karena disebabkan oleh ketidaksesuaian dengan indikator, redaksi soal yang kurang jelas, tingkat kesukaran yang tidak proporsional, atau distraktor yang tidak berfungsi dengan baik.(Ida & Musyarofah, 2021) Sementara itu, nilai Cronbach's Alpha sebesar 0,716 menunjukkan bahwa instrumen memiliki konsistensi internal yang baik dan reliabel, yang dipengaruhi oleh keseragaman konstruk, jumlah item yang cukup, serta dominasi butir soal yang valid dan memiliki daya pembeda yang baik, sehingga secara umum instrumen

layak digunakan meskipun masih perlu perbaikan pada beberapa butir soal.(Magdalena et al., 2021)

Hasil analisis tingkat kesukaran menunjukkan bahwa sebagian besar butir soal berada pada kategori mudah, sehingga instrumen yang digunakan belum sepenuhnya ideal. Secara teoritis, soal yang baik seharusnya didominasi oleh kategori sedang agar mampu mengukur kemampuan mahasiswa secara optimal.(Utami, 2023) Dominasi soal mudah dalam penelitian ini mengindikasikan bahwa sebagian besar mahasiswa dapat menjawab soal dengan benar, sehingga instrumen kurang mampu memberikan tantangan yang cukup dalam mengukur variasi kemampuan mahasiswa.(Yustiandi, 2024) Ditinjau dari aspek daya pembeda, hasil penelitian menunjukkan bahwa sebagian besar soal memiliki daya pembeda yang baik hingga sangat baik. Hal ini berarti bahwa soal mampu membedakan mahasiswa berkemampuan tinggi dan rendah secara efektif. Dengan demikian, meskipun tingkat kesukaran soal cenderung mudah, kualitas soal dari segi kemampuan diskriminatif masih tergolong baik dan dapat digunakan sebagai alat evaluasi pembelajaran.

Sementara itu, dari aspek efektivitas distraktor, secara umum pengecoh dalam soal telah berfungsi dengan cukup baik, terutama pada soal yang memiliki daya pembeda tinggi. Namun, pada beberapa soal yang tergolong mudah, distraktor cenderung kurang efektif karena tidak mampu menarik perhatian mahasiswa, sehingga opsi jawaban yang salah mudah dieliminasi. Hal ini menunjukkan bahwa kualitas distraktor masih perlu ditingkatkan agar lebih homogen dan sesuai dengan karakteristik kesalahan umum mahasiswa. Secara teoritis, distraktor yang baik adalah distraktor yang plausibel (masuk akal), homogen dengan kunci jawaban, memiliki panjang dan struktur yang relatif sama, serta mampu dipilih oleh sebagian peserta didik berkemampuan rendah, sehingga berfungsi untuk mengecoh dan meningkatkan daya pembeda soal. Selain itu, distraktor yang baik tidak boleh terlalu jelas salah atau menyimpang dari konteks soal, karena hal tersebut akan menyebabkan peserta didik dengan mudah menebak jawaban yang benar.(Rahmawati & Rahman, 2025) Penyusunan distraktor perlu mempertimbangkan pola kesalahan umum mahasiswa agar opsi jawaban yang disediakan benar-benar mampu menguji pemahaman secara mendalam.

Hasil penelitian ini memiliki implikasi penting terhadap pembelajaran Bahasa Arab, khususnya dalam aspek pemahaman teks, qawā‘id, dan mufradāt. Instrumen evaluasi yang kurang bervariasi tingkat kesukarannya berpotensi hanya mengukur kemampuan dasar mahasiswa, seperti mengingat dan memahami, tanpa mampu mengungkap kemampuan berpikir tingkat tinggi.(M Baihaqi, A Syarifah, 2025) Oleh karena itu, diperlukan penyusunan soal yang lebih beragam agar dapat mengukur kemampuan analisis dan pemahaman mendalam terhadap struktur bahasa Arab.

Temuan penelitian ini menguatkan hasil penelitian terdahulu yang menegaskan bahwa analisis butir soal merupakan tahapan fundamental dalam meningkatkan kualitas instrumen evaluasi pembelajaran. Analisis ini tidak hanya berfungsi untuk mengidentifikasi validitas dan reliabilitas instrumen, tetapi juga untuk menilai ciri khusus pada tiap butir soal melalui indikator tingkat kesukaran, daya pembeda, dan efektivitas distraktor. Melalui analisis tersebut, pendidik dapat memperoleh gambaran empiris mengenai sejauh mana instrumen mampu mengukur kemampuan peserta didik secara akurat dan objektif.(Nurhayati & Sokosari, 2024) Namun demikian, Hasil penelitian ini menunjukkan perbedaan dengan beberapa penelitian sebelumnya yang umumnya menghasilkan distribusi tingkat kesukaran yang lebih seimbang, yaitu didominasi oleh soal kategori sedang. Dalam penelitian ini, distribusi soal masih didominasi oleh kategori mudah, yang mengindikasikan bahwa instrumen cenderung belum mampu

mengakomodasi variasi kemampuan peserta didik secara optimal, terutama dalam mengukur kemampuan berpikir tingkat tinggi. Kondisi ini berimplikasi pada rendahnya sensitivitas instrumen dalam membedakan tingkat kemampuan mahasiswa secara lebih mendalam.

Secara praktis, hasil penelitian ini memberikan rekomendasi bagi dosen atau pengajar Bahasa Arab untuk melakukan revisi terhadap butir soal yang terlalu mudah, mengembangkan bank soal berbasis analisis empiris, serta menerapkan analisis butir soal secara rutin setelah pelaksanaan tes. Selain itu, penggunaan perangkat lunak seperti Microsoft Excel atau SPSS juga dapat membantu dalam melakukan analisis secara lebih sistematis dan akurat.(Sudrajat, 2025)

Penelitian ini memiliki keterbatasan, di antaranya jumlah sampel yang relatif kecil dan hanya melibatkan satu kelas, sehingga hasil penelitian belum dapat digeneralisasikan secara luas. Selain itu, data yang digunakan terbatas pada hasil tes pilihan ganda tanpa analisis mendalam terhadap pilihan distraktor secara spesifik. Berdasarkan hasil penelitian, disarankan agar penelitian selanjutnya menggunakan jumlah sampel yang lebih besar serta mengembangkan instrumen soal yang mampu mengukur kemampuan berpikir tingkat tinggi (HOTS). Selain itu, analisis butir soal sebaiknya dilakukan secara berkelanjutan dengan memanfaatkan teknologi untuk meningkatkan kualitas evaluasi pembelajaran Bahasa Arab secara lebih komprehensif.(Pahlefi, 2022)

KESIMPULAN

Penelitian ini menunjukkan bahwa Analisis butir soal merupakan tahapan penting dalam menjamin kualitas instrumen evaluasi pembelajaran Bahasa Arab, khususnya dalam mengukur pemahaman mahasiswa pada aspek qawā‘id. Sebagaimana diharapkan dalam tujuan penelitian, instrumen yang dianalisis diharapkan mampu mengukur kemampuan mahasiswa secara akurat melalui tingkat kesukaran, daya pembeda, dan efektivitas distraktor. Hasil penelitian menunjukkan bahwa meskipun sebagian besar butir soal memiliki daya pembeda yang baik sehingga mampu mengidentifikasi perbedaan kemampuan mahasiswa, distribusi tingkat kesukaran masih belum ideal karena didominasi oleh soal kategori mudah. Hal ini mengindikasikan bahwa instrumen belum sepenuhnya mampu mengukur variasi kemampuan mahasiswa secara komprehensif, khususnya pada tingkat berpikir yang lebih tinggi. Selain itu, efektivitas distraktor secara umum telah menunjukkan fungsi yang cukup baik, namun masih terdapat beberapa butir soal yang perlu disempurnakan agar lebih mampu mengecoh mahasiswa berkemampuan rendah. Temuan ini menegaskan bahwa kualitas instrumen evaluasi tidak hanya ditentukan oleh ketepatan materi, tetapi juga oleh keseimbangan tingkat kesukaran dan kualitas pengecoh yang digunakan. Dengan demikian, penelitian ini memberikan makna bahwa analisis butir soal tidak hanya berfungsi sebagai evaluasi hasil tes, tetapi juga sebagai dasar dalam memperbaiki desain pembelajaran dan penyusunan instrumen yang lebih berkualitas. Ke depan, hasil penelitian ini membuka peluang untuk pengembangan instrumen evaluasi Bahasa Arab yang lebih komprehensif dengan mengintegrasikan soal berbasis kemampuan berpikir tingkat tinggi (HOTS) serta pemanfaatan teknologi dalam analisis data. Penelitian selanjutnya disarankan untuk melibatkan jumlah sampel yang lebih luas dan beragam, serta mengkaji efektivitas instrumen pada berbagai konteks materi Bahasa Arab lainnya.

UCAPAN TERIMA KASIH

Penulis menyampaikan apresiasi kepada Fakultas Tarbiyah dan Ilmu Keguruan UIN Sunan Ampel Surabaya serta mahasiswa program intensif Bahasa Arab yang telah berpartisipasi dalam penelitian ini. Ucapan terima kasih juga ditujukan kepada seluruh pihak yang telah memberikan dukungan, bantuan, dan kontribusi selama proses penelitian hingga penyusunan artikel. Diharapkan penelitian ini dapat memberikan manfaat bagi pengembangan evaluasi pembelajaran Bahasa Arab.

DAFTAR PUSTAKA

Alfinnas, H. (2026). Instrumen Penilaian Autentik Pada Pembelajaran Bahasa Arab Di Madrasah Aliyah. *Aphorisme: Journal of Arabic Language, Literature, and Education*, *7*(1), 193–210. https://doi.org/10.37680/aphorisme.v7i1.9131

Asria, L. (2023). Analisis Butir Soal Penilaian Tengah Semester (PTS) Matematika Kelas XI Berdasarkan Teori Klasik Item Analysis of Mathematics Mid-Semester Assessment for Class XI Based on Classical Theory. *MATH LOCUS: Jurnal Riset Dan Inovasi Pendidikan Matematika*, *4*(1), 1–11. https://doi.org/https://doi.org/10.31002/mathlocus.v4i1.3177

Harsono, I., Abidin, M., & Hilmi, D. (2024). Analisis Butir Soal Bahasa Arab Di SDIT Al Islam Kudus menggunakan Metode Distinguishing , Difficulty , dan Dispersion. *Al-Mu'arrib: Jurnal Pendidikan Bahasa Arab*, *4*(2), 109–120. https://doi.org/DOI: https:// 10.32923/al-muarrib.v4i2.4672

Hartati, N., Pratama, H., Yogi, S., & Prof, U. M. (2019). Item Analysis for a Better Quality Test. *ELIF*, *2*(1), 59–70. https://doi.org/https://jurnal.umj.ac.id/index.php/ELIF

Ida, F. F., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *Al-Mu'Arrib: Journal of Arabic Education*, *1*(1), 34–44. https://doi.org/10.32923/al-muarrib.v1i1.2100

M Baihaqi, A Syarifah, M. A. (2025). Enhancing Arabic Translation Competency In Higher Education: An Evaluation Of The Becoming A Translation Practitioner'Program. *Ijaz Arabi Journal of Arabic Learning. https://doi.org/10.18860/ijazarabi.v8i1.31642*.

Magdalena, I., Fauziah, S. N., Faziah, S. N., & Nupus, F. S. (2021). Analisis Validitas, Reliabilitas, Tingkat Kesulitan Dan Daya Beda Butir Soal Ujian Akhir Semester Tema 7 Kelas Iii Sdn Karet 1 Sepatan. *BINTANG : Jurnal Pendidikan Dan Sains*, *3*(2), 198–214. https://ejournal.stitpn.ac.id/index.php/bintang

Maria, T., Juniyati, A., Ubaid, S., Islam, U., Syarif, N., & Jakarta, H. (2025). Penilaian Ranah Kognitif dalam Pembelajaran Bahasa Arab : Pembelajaran bahasa Arab di lembaga pendidikan memiliki tujuan yang tidak hanya berfokus pada penguasaan keterampilan berbahasa , tetapi memahami struktur bahasa , makna , dan konteks penggunaannya. *Qawaid*, *1*(02), 55–68.

Nadapdap, M. J., Yasmin, P., Purba, S., Degenerative, P. U. I. P., Medicine, L., & Indonesia, U. P. (2025). Pengaruh Perilaku Mahasiswa Terhadap Proses ARDA dengan Hasil Ujian CBT S1 Kedokteran FK UNPRI Tahun 2025. *JURNAL LOCUS: Penelitian & Pengabdian*, *4*(11), 10949–10967. https://doi.org/10.58344/locus.v4i11.5124.

Naja, A. F., Arlisyah, M., & Utami, P. (2025). Pengembangan Instrumen Diagnostik Berformat

Four Tier Multiple Choice pada Materi Operasi Aljabar. *Suska Journal of Mathematics Education*, *11*(1), 9–22. https://doi.org/10.24014/sjme.v11i1.31306.

Nurhayati, I., & Sokosari, I. (2024). Pengembangan Alat Evaluasi Pembelajaran Dengan Aplikasi Gimkit Pada Mata Pelajaran Ipas Siswa Madrasah Ibtidaiyah. *Al-Adawat*, *03*(01), 66–80. https://doi.org/10.33752/aldawat.v3i01.7075.

Pahlefi, M. R. (2022). Pengembangan Instrumen Penilaian Keterampilan Menyimak (Mahārah al-Istima') dalam Pembelajaran Bahasa Arab. *Uktub: Journal of Arabic Studies*, *2*(2), 68–84. https://doi.org/10.32678/uktub.v2i2.6458

Radha, L. (2025). Analisis butir soal pengukuran kemampuan berpikir kritis materi pencemaran lingkungan melalui teori tes klasik dan Rasch model. *Bio-Pedagogi : Jurnal Pembelajaran Biologi*, *14*(1), 36–46. . https://dx.doi.org/10.20961/bio-pedagogi.v14i1.88792.

Rahmawati, N. S., & Rahman, M. F. (2025). Analisis Butir Soal Objektif Penilaian Sumatif Akhir Semester Gasal Mata Pelajaran Ekonomi. *JIIP (Jurnal Ilmiah Ilmu Pendidikan)*, *8*, 8863–8873.

Said, A. S., Fikri, M., Fadjri, N., Negeri, I., Kalijaga, S., Yogyakarta, M., Info, A., & History, A. (2025). Pengaruh Bi'ah Lughowiyyah terhadap Maharah Kalām dalam Pembelajaran Bahasa Arab. *Mahira: Journal of Arabic Studies and Teaching*, *3*(3), 155–165. https://doi.org/DOI: https://doi.org/10.14421/mahira.2025.33.01

Saputra, H. D., Purwanto, W., Setiawan, D., Fernandez, D., & Putra, R. (2022). Hasil Belajar Mahasiswa: Analisis Butir Soal Tes. *Edukasi: Jurnal Pendidikan*, *20*(1), 15–27. https://doi.org/10.31571/edukasi.v20i1.3432

Sudrajat, A. E. (2025). Instrumen Evaluasi Maharah Kalam Dalam Pembelajaran Bahasa Arab. *Al-Tadris: Jurnal Pendidikan Bahasa Arab*, 615–641.

Utami, Y. (2023). Uji Validitas dan Uji Reliabilitas Instrument Penilaian Kinerja Dosen. *Jurnal Sains Dan Teknologi*, *4*(2), 21–24. https://doi.org/10.55338/saintek.v4i2.730

Yusron, E., Retnawati, H., & Rafi, I. (2020). Bagaimana hasil penyetaraan paket tes USBN pada mata pelajaran matematika dengan teori respons butir ? *Jurnal Riset Pendidikan Matematika*, *7*(1), 1–12. https://doi.org/10.21831/jrpm.v7i1.31221.

Yustiandi. (2024). Analisis Karakteristik Butir Soal Seleksi Penerimaan Siswa Baru Berdasarkan Pendekatan Teori Tes Klasik. *Jurnal Basicedu*, *8*(6), 4700–4706. https://doi.org/10.31004/basicedu.v8i6.9031.