Vol. 21, No. 2, November 2025 *p-ISSN: 2087-3476*
Jurnal Pepatuzdu: Media Pendidikan dan Sosial Kemasyarakat *e-ISSN: 2541-5700*
DOI: http://dx.doi.org/10.35329/fkip.v21i2.6040

ANALISIS KUALITAS TES BAHASA ARAB DI INDONESIA: STUDI SYSTEMATIC LITERATURE REVIEW TENTANG VALIDITAS, RELIABILITAS, TINGKAT KESUKARAN, DAN DAYA BEDA

Ulya Nur Alim[1], Syahrul[1], Luís Miguel Oliveira de Barros Cardoso[2], Suryadi Ishak[1*], Andi Asrifan[1]

[1]Universitas Negeri Makassar
[2]Polytechnic Institute of Portalegre
***Email: suryadi.ishak@unm.ac.id**

ABSTRACT

Evaluation in Arabic language learning is essential to measure students' achievement; however, the quality of tests used in Indonesia still requires improvement. This study employed the Systematic Literature Review (SLR) method to analyze the validity, reliability, difficulty level, and discrimination power of Arabic test items, based on a synthesis of six articles indexed in SINTA and Scopus, published between 2019 and 2024. This SLR approach offers a new contribution by systematically revealing national trends and gaps in item quality, which have not been comprehensively analyzed in previous studies. The findings show that on average, 66% of the items were valid, and most tests demonstrated very high reliability (≥ 0.85), although some tests had low reliability (0.54). The distribution of difficulty levels was imbalanced, with 50.83% of items being too easy and only 7.67% classified as difficult, deviating from the ideal distribution. Additionally, 34% of the items exhibited low discrimination power, reducing the effectiveness of assessments in distinguishing students' abilities. These imbalances can lead to biased evaluations and hinder students' competency development. The practical implications of this study include the importance of teacher training in item analysis, the application of Bloom's Taxonomy to balance item difficulty levels, and the development of a standardized, data-driven item bank. The main contribution of this research is to provide empirical foundations for improving Arabic language assessment policies in Indonesia and to propose a more accurate and fair evidence-based evaluation approach.

Keywords: Arabic test, difficulty level, item discrimination, reliability, validity

PENDAHULUAN

Kualitas tes dalam pembelajaran Bahasa Arab sangat ditentukan oleh karakteristik butir soal yang menyusunnya, seperti validitas, reliabilitas, tingkat kesukaran, dan daya beda. Tes yang baik bukan hanya sekadar alat ukur pencapaian pembelajaran, tetapi juga berfungsi sebagai landasan dalam pengambilan keputusan pendidikan, baik di tingkat kelas maupun institusi. Tes yang disusun tanpa memperhatikan karakteristik tersebut berpotensi menghasilkan informasi yang keliru mengenai kemampuan peserta didik, dan pada akhirnya memengaruhi mutu proses pembelajaran itu sendiri.

Namun, berbagai penelitian menunjukkan bahwa guru Bahasa Arab di Indonesia masih jarang melakukan analisis butir soal secara sistematis. Minimnya pelatihan, kurangnya pemahaman tentang teknik evaluasi, serta keterbatasan akses terhadap perangkat bantu seperti *software Iteman*, menjadi sejumlah faktor yang menghambat implementasi evaluasi yang akurat dan objektif (Ismiyati dkk., 2023; Musa dkk., 2024). Akibatnya, banyak soal yang digunakan dalam asesmen pembelajaran tidak melalui proses validasi dan analisis karakteristik, sehingga tidak dapat secara optimal mengukur kompetensi peserta didik secara menyeluruh.

Minimnya pelatihan dan pemahaman guru terhadap analisis butir soal berdampak serius terhadap kualitas instrumen evaluasi pembelajaran, khususnya pada mata pelajaran Bahasa Arab. Hal ini berpotensi menghasilkan keputusan yang tidak akurat dalam menilai capaian belajar siswa, serta menghambat pengembangan sistem evaluasi yang objektif dan adil.

Berdasarkan pengalaman penulis saat mengajar dan melakukan penelitian di beberapa sekolah, ditemukan bahwa praktik evaluasi oleh guru sering kali tidak didasarkan pada pengukuran objektif terhadap kemampuan siswa. Banyak guru yang lebih fokus pada pencapaian target nilai institusional, sehingga soal evaluasi dibuat terlalu mudah, tidak divalidasi, dan tidak dianalisis kualitasnya. Dalam beberapa kasus, nilai diberikan hanya berdasarkan kehadiran atau sikap siswa, tanpa dilakukan remedial atau asesmen ulang meskipun hasil belajar belum memenuhi standar. Praktik ini menghasilkan data nilai dalam rapor yang tidak mencerminkan kemampuan akademik siswa yang sebenarnya. Ketidaksesuaian antara nilai dan kompetensi riil ini berkontribusi terhadap rendahnya hasil asesmen internasional seperti PISA, yang menilai langsung kemampuan membaca, matematika, dan sains secara objektif dan independen (Budiono & Hatip, 2023). Fenomena tersebut menunjukkan adanya kesenjangan serius antara penilaian sekolah dan penguasaan kompetensi dasar siswa Indonesia, yang memperkuat urgensi perlunya peningkatan kualitas instrumen evaluasi di sekolah, termasuk dalam mata pelajaran Bahasa Arab.

Meskipun sejumlah penelitian telah dilakukan untuk menganalisis karakteristik butir soal secara individual, belum ada kajian sistematis yang secara menyeluruh memetakan tren kualitas soal Bahasa Arab di Indonesia berdasarkan publikasi nasional dalam enam tahun terakhir. Oleh karena itu, penelitian ini hadir untuk mengisi kesenjangan tersebut dengan memberikan sintesis yang komprehensif terhadap hasil-hasil penelitian sebelumnya.

Istilah "alat" sering juga disebut sebagai "instrumen", sehingga alat evaluasi dapat dipahami sebagai instrumen evaluasi. Instrumen yang baik adalah instrumen yang mampu mengukur objek evaluasi secara tepat sesuai dengan kondisi sebenarnya. Instrumen evaluasi terbagi dua yaitu instrumen tes dan non tes (Miladya, 2021). Dalam pembelajaran Bahasa Arab, evaluasi yang efektif memerlukan salah satunya yaitu instrumen tes dengan karakteristik teknis yang memadai. Empat indikator utama yang menjadi acuan kualitas tes meliputi validitas, reliabilitas, tingkat kesukaran, dan daya pembeda (Arifianto dkk., 2021). Keempat aspek ini tidak hanya menentukan kualitas tes secara statistik, tetapi juga mencerminkan keadilan dan ketepatan dalam mengukur kemampuan peserta didik. Dengan demikian, instrumen yang memenuhi kriteria tersebut dapat dijadikan dasar yang terpercaya untuk menilai hasil belajar sekaligus memberikan umpan balik yang konstruktif bagi proses pembelajaran.

Validitas merujuk pada sejauh mana suatu instrumen benar-benar mengukur apa yang seharusnya diukur (Muzaffar, 2016). Dalam konteks pembelajaran Bahasa Arab, validitas berkaitan dengan kesesuaian antara isi soal dan kompetensi linguistik yang ditargetkan. Tes yang valid memastikan bahwa hasil yang diperoleh mencerminkan kompetensi yang sebenarnya dimiliki peserta didik, bukan kemampuan di luar domain yang diukur.

Reliabilitas menggambarkan konsistensi hasil yang dihasilkan oleh suatu instrumen tes. Tes yang reliabel akan memberikan hasil yang relatif sama apabila diberikan kepada kelompok peserta didik yang sama dalam kondisi yang sebanding (Alwinda, 2020).

Tingkat kesukaran soal menunjukkan proporsi peserta didik yang dapat menjawab suatu soal dengan benar. Soal yang terlalu mudah atau terlalu sulit cenderung menurunkan efektivitas evaluasi karena tidak mampu menangkap variasi kemampuan siswa secara optimal. Soal dengan tingkat kesukaran sedang dinilai paling efektif karena dapat memfasilitasi pemetaan kemampuan siswa secara lebih akurat (Zainuddin & Fatmawati, 2022).

Daya beda menunjukkan kemampuan suatu butir soal untuk membedakan antara peserta didik yang berkemampuan tinggi dan rendah (Fatimah & Alfath, 2019). Soal dengan daya beda tinggi memungkinkan pengambilan keputusan pendidikan yang lebih tajam, sedangkan daya beda rendah menunjukkan soal yang kurang bermanfaat untuk asesmen sumatif.

Interaksi antara keempat aspek ini membentuk dasar penilaian kualitas tes. Suatu tes tidak dapat dikatakan berkualitas apabila hanya valid tetapi tidak reliabel, atau reliabel tetapi memiliki distribusi tingkat kesukaran yang tidak proporsional dan daya beda yang rendah. Oleh karena itu, analisis butir soal secara menyeluruh menjadi penting agar asesmen benar-benar mampu mencerminkan kompetensi peserta didik secara objektif.

Berdasarkan fenomena tersebut, muncul pertanyaan: Apakah butir soal Bahasa Arab yang digunakan di Indonesia telah memenuhi standar kualitas yang ideal, khususnya dari aspek validitas, reliabilitas, daya beda, dan tingkat kesukaran?

Penelitian ini bertujuan untuk menganalisis tren kualitas soal Bahasa Arab di Indonesia melalui kajian sistematis terhadap artikel-artikel yang dipublikasikan antara tahun 2019 hingga 2024. Fokus utama terletak pada validitas, reliabilitas, tingkat kesukaran, dan daya beda soal, sebagai indikator utama dalam evaluasi instrumen tes.

Penelitian ini berhipotesis bahwa masih terdapat banyak soal Bahasa Arab yang belum memenuhi standar kualitas ideal, terutama karena kurangnya praktik analisis butir soal oleh guru secara sistematis di lapangan.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan *Systematic Literature Review* (SLR) untuk mengkaji kualitas butir soal dalam tes Bahasa Arab di Indonesia. Fokus kajian diarahkan pada empat aspek utama yang menjadi indikator kualitas tes, yaitu validitas, reliabilitas, tingkat kesukaran, dan daya beda. Pendekatan SLR dipilih karena memungkinkan peneliti untuk menyusun sintesis menyeluruh dari hasil-hasil penelitian terdahulu, sehingga dapat memberikan pemahaman yang lebih luas dan mendalam mengenai tren dan kesenjangan dalam penyusunan tes Bahasa Arab.

Proses pengumpulan data dilakukan melalui pencarian literatur dari tiga basis data utama, yaitu Google Scholar, SINTA, dan Scopus. Kata kunci yang digunakan dalam proses pencarian meliputi: *validitas, reliabilitas, tingkat kesukaran, daya beda butir soal Bahasa Arab, analisis butir soal Bahasa Arab, dan evaluasi pembelajaran Bahasa Arab*. Dari hasil pencarian awal, peneliti berhasil mengumpulkan sejumlah artikel tanpa pembatasan tahun publikasi. Namun, untuk menjamin kualitas dan relevansi, kemudian diterapkan kriteria inklusi dan eksklusi.

Kriteria inklusi dalam penelitian ini mencakup artikel yang: (1) diterbitkan dalam enam tahun terakhir (2019–2024), (2) dimuat dalam jurnal ilmiah bereputasi yang terindeks SINTA atau Scopus, dan (3) secara substantif membahas karakteristik butir soal yang menjadi fokus penelitian (validitas, reliabilitas, daya beda, dan tingkat kesukaran soal). Artikel yang hanya

bersifat konseptual tanpa analisis data empiris tidak dimasukkan dalam kajian. Selain itu, hanya artikel ilmiah berbentuk jurnal *peer-reviewed* yang dijadikan sumber data, sementara skripsi, tesis, disertasi, prosiding, maupun laporan institusional dikecualikan.

Kriteria eksklusi diterapkan pada artikel-artikel yang tidak memenuhi syarat di atas, termasuk yang terbit sebelum tahun 2019, tidak berasal dari jurnal terindeks, atau tidak memuat pembahasan yang relevan terkait karakteristik butir soal. Artikel yang hanya membahas teori atau pendekatan pembelajaran Bahasa Arab tanpa mengaitkannya dengan data kuantitatif mengenai soal evaluasi juga tidak dimasukkan dalam analisis.

Selain artikel-artikel yang dianalisis secara sistematis, penelitian ini juga mengacu pada beberapa sumber teoritis tambahan yang digunakan untuk memperkuat landasan konseptual dan pembahasan. Artikel-artikel tersebut tidak termasuk dalam analisis utama karena tidak memuat data kuantitatif terkait butir soal, namun tetap relevan dalam menyusun kerangka berpikir penelitian.

Setelah dilakukan penyaringan berdasarkan kriteria inklusi dan eksklusi, dari total 30 artikel yang berhasil dikumpulkan, hanya enam artikel yang memenuhi seluruh kriteria dan dijadikan sebagai sumber data utama dalam analisis sistematis. Keenam artikel ini dipilih karena secara substantif membahas aspek-aspek kualitas butir soal, khususnya validitas, reliabilitas, tingkat kesukaran, dan daya beda, serta menyajikan data kuantitatif yang relevan. Setiap artikel dikaji secara mendalam untuk mengidentifikasi temuan-temuan dalam keempat aspek tersebut, kemudian disusun secara sistematis dan diklasifikasikan untuk dianalisis secara deskriptif. Analisis ini bertujuan untuk mengungkap pola-pola umum, perbedaan hasil, serta potensi kesenjangan dalam penelitian-penelitian sebelumnya terkait kualitas tes Bahasa Arab di Indonesia.

Selain artikel-artikel yang dijadikan data utama dalam analisis, beberapa artikel lainnya dari 30 bacaan awal tetap digunakan sebagai referensi konseptual dalam penyusunan kerangka teori dan pembahasan. Meskipun artikel-artikel ini tidak dimasukkan ke dalam tahap sintesis data karena tidak menyajikan hasil analisis butir soal secara kuantitatif, keberadaannya tetap relevan dalam memperkuat dasar teori dan konteks akademik penelitian ini.

Perlu dicatat bahwa seluruh proses seleksi dan analisis dilakukan secara mandiri oleh peneliti tanpa keterlibatan reviewer lain dalam validasi data. Oleh karena itu, hasil kajian ini tetap memiliki keterbatasan dari sisi potensi subjektivitas penilaian. Meskipun demikian, prosedur telah dirancang seobjektif mungkin dengan berpedoman pada prinsip-prinsip transparansi dan sistematika dalam metode SLR.

Vol. 21, No. 2, November 2025 *p-ISSN: 2087-3476*
Jurnal Pepatuzdu: Media Pendidikan dan Sosial Kemasyarakat *e-ISSN: 2541-5700*
DOI: http://dx.doi.org/10.35329/fkip.v21i2.6040

HASIL DAN PEMBAHASAN

Ringkasan Temuan Literatur

Analisis sistematis terhadap kualitas butir soal bahasa Arab di sekolah-sekolah Indonesia dilakukan untuk menjawab rumusan masalah dalam penelitian ini dengan meninjau validitas, reliabilitas, daya beda, dan tingkat kesukaran soal berdasarkan berbagai studi yang telah dilakukan sebelumnya. Studi-studi ini memberikan gambaran mengenai kualitas butir soal yang digunakan dalam evaluasi pembelajaran bahasa Arab serta mengidentifikasi kelemahan dan aspek yang perlu diperbaiki. Tabel berikut menyajikan ringkasan temuan utama dari berbagai penelitian yang telah dianalisis.

Tabel 1. Hasil penelitian-penelitian kualitas dan karakter butir soal bahasa Arab

No.	Author	Tahun	Judul Penelitian	Sampel	Hasil Penelitian
1	Muhammad Taufiq Ismail & Farikh Marzuki Ammar	2024	Analisis Butir Soal Pelajaran Bahasa Arab Sumatif Akhir Semester Ganjil Tahun Ajaran 2022/2023	20 siswa kelas XI SMA Al-Fattah Sidoarjo	14 soal valid (56%), 11 soal tidak valid (44%). Reliabilitas baik. Tingkat kesukaran: 11 sedang (44%), 10 mudah (40%), 4 sangat mudah (16%).
2	Siti Fathimah Al Fathiyah	2019	Analisis Butir Soal Pelajaran Bahasa Arab Di MA Roudlotul Ulum Pagak Malang	13 siswa kelas XI-IPS di MA Roudlotul Ulum Pagak Malang	3 soal valid (2.85%), 12 soal tidak valid (97.15%). Reliabilitas sedang (0,54). Tingkat kesukaran: 8 soal mudah, 3 sedang, 4 sukar. Daya beda: 3 soal baik, 3 sedang, sisanya tidak bisa membedakan

					siswa.
3	Bahrudin Fahmi, Syahrul Rizqi, Nurul Elmira H	2022	Analisis Butir Soal Bahasa Arab Siswa MAS Pondok Pesantren Asaalam Kampar Riau	32 siswa kelas XI MAS Pondok Pesantren Asaalam	Daya beda: 7 soal sangat tinggi, 9 tinggi, 10 sedang, 13 rendah, 11 sangat rendah. Tingkat kesukaran: 43 soal mudah, 7 sedang. Validitas: 24 soal valid (48%), 26 tidak valid (52%). Reliabilitas sangat tinggi (0,868).
4	Nurul Fikriyah	2021	Analisis Butir Soal Ulangan Tengah Semester Mata Pelajaran Bahasa Arab	115 siswa kelas VII SMP Muhammadiyah 1 Yogyakarta	31 soal valid (89%), 4 tidak valid (11%). Reliabilitas tinggi (0,854). Tingkat kesukaran: 3 soal mudah (10%), 25 sedang (80%), 3 sulit (10%). Daya beda: 12 baik (39%), 13 cukup (42%), 6 jelek (19%).
5	Deni Maulana, Anwar Sanusi	2020	Analisis Butir Soal Bahasa Arab Ujian Akhir Madrasah Bersama Daerah (UAMBD)	27 siswa Madrasah Ibtidaiyah (MI) Tahun Ajaran 2017-2018	Validitas sangat tinggi (100% sesuai kisi-kisi). Reliabilitas tinggi (0,68). Tingkat kesukaran: 63,23% dapat digunakan

			Madrasah Ibtidaiyah Tahun 2017-2018		sebagai tes standar. Daya beda: 31,4% baik, 25,7% perlu diperbaiki.
6	Rahmat Danni, Ajeng Wahyuni, Tauratiya	2021	Item Response Theory Approach: Kalibrasi Butir Soal Penilaian Akhir Semester Mata Pelajaran Bahasa Arab	176 siswa kelas XI MAN 1 Pangkalpinang Tahun Ajaran 2019/2020	40 soal valid (100%). Reliabilitas sangat tinggi (0,884). Tingkat kesulitan dan daya beda baik pada 33 soal (82,5%), 7 soal (17,5%) perlu revisi/eliminasi.

(Sumber: Hasil Analisis Data, 2025)

Analisis Kualitas dan Karakteristik Butir Soal

Validitas

Berikut merupakan tabel persentase soal valid dan tidak valid dari berbagai penelitian.

Tabel 2. Persentase butir soal dengan validitas baik

No.	Author	Tahun	Jumlah Soal	Persentase Validitas
1	Muhammad Taufiq Ismail & Farikh Marzuki Ammar	2024	14 dari 25 butir soal	56%
2	Siti Fathimah Al Fathiyah	2019	3 dari 15 butir	2.85%
3	Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H	2022	24 dari 50 butir soal	48%
4	Nurul Fikriyah	2021	31 dari 35 butir soal	89%
5	Deni Maulana & Anwar Sanusi	2020	35 dari 35 butir soal	100%
6	Rahmat Danni, Ajeng Wahyuni, Tauratiya	2021	40 dari 40 butir soal	100%
Rata-rata persentase				66%

(Sumber: Hasil Analisis Data, 2025)

Berdasarkan pembahasan yang diperoleh dari beberapa penelitian diatas diketahui bahwa mayoritas hasil penelitian dari berbagai studi menunjukkan butir soal yang valid, namun masih banyak ditemukan butir soal yang tidak valid. Persentase butir soal yang valid dikemukakan pada studi berikut: Muhammad Taufiq Ismail & Farikh Marzuki Ammar (2024) menemukan sebanyak 14 dari 25 (56%) butir soal yang valid, Siti Fathimah Al Fathiyah (2019) menemukan butir soal yang valid hanya 3 dari 15 butir (2.85%), Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H (2022) menemukan 24 dari 50 (48%) soal yang berstatus valid, Nurul Fikriyah (2021) menemukan kategori butir soal valid berjumlah 31 dari 35 butir soal (89%), Deni Maulana & Anwar Sanusi (2020) menemukan 35 dari 35 (100%) butir soal valid, Terakhir Rahmat Danni, Ajeng Wahyuni, Tauratiya (2021) menemukan 40 dari 40 (100%) butir soal valid. Jika hasil dari seluruh studi tersebut disimpulkan maka persentase rata-rata soal yang valid berjumlah 66%.

Tabel 3. Persentase butir soal tidak valid

No.	Author	Tahun	Jumlah Soal	Persentase Tidak Valid
1	Muhammad Taufiq Ismail & Farikh Marzuki Ammar	2024	11 dari 25 butir soal	44%
2	Siti Fathimah Al Fathiyah	2019	12 dari 15 butir soal	97.15%
3	Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H	2022	26 dari 50 butir soal	52%
4	Nurul Fikriyah	2021	4 dari 35 butir soal	11%
5	Deni Maulana & Anwar Sanusi	2020	0 dari 35 butir soal	0%
6	Rahmat Danni, Ajeng Wahyuni, Tauratiya	2021	0 dari 40 butir soal	0%
Rata-rata persentase				34%

(Sumber: Hasil Analisis Data, 2025)

Beberapa soal yang tidak valid dari berbagai studi yang telah dipaparkan dirincikan lebih lanjut sebagai berikut: Muhammad Taufiq Ismail & Farikh Marzuki Ammar (2024) menemukan sebanyak 11 dari 25 (44%) butir soal yang tidak valid, Siti Fathimah Al Fathiyah (2019) menemukan kategori soal tidak valid 12 dari 15 butir soal (97.15%), Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H (2022) menemukan 26 dari 50 (52%) soal yang berstatus tidak valid, Nurul Fikriyah (2021) menemukan soal yang termasuk ke dalam kategori tidak valid berjumlah 4 dari 35 butir soal (11%), Deni Maulana & Anwar Sanusi (2020)

tidak menemukan butir soal tidak valid (0%), Terakhir Rahmat Danni, Ajeng Wahyuni, Tauratiya (2021) tidak menemukan butir soal tidak valid (0%). Jika hasil dari seluruh studi tersebut disimpulkan maka persentase rata-rata soal yang tidak valid berjumlah 34%.

Reliabilitas Tes

Sharma (2016) memaparkan pada studinya mengenai kriteria untuk menilai koefisien reliabilitas. Kriteria reliabilitas tersebut disajikan pada tabel berikut.

Tabel 4. Pedoman kriteria reliabilitas

Kriteria Reliabilitas	Kategori
≥ 0.80	Sangat tinggi
0.70 - 0.79	Tinggi
0.60 - 0.69	Cukup
0.40 - 0.59	Rendah
< 0.40	Sangat rendah

(Sumber: Sharma, 2016)

Tabel 5. Analisis reliabilitas soal

No	Author	Tahun	Nilai Reliabilitas	Kategori
1	Muhammad Taufiq Ismail & Farikh Marzuki Ammar	2024	0,861	sangat tinggi
2	Siti Fathimah Al Fathiyah	2019	0,54	Rendah
3	Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H	2022	0,868	sangat tinggi
4	Nurul Fikriyah	2021	0,854	sangat tinggi
5	Deni Maulana & Anwar Sanusi	2020	0,68	Cukup
6	Rahmat Danni, Ajeng Wahyuni, Tauratiya	2021	0,884	sangat tinggi
Kategori				sangat tinggi

(Sumber: Hasil Analisis Data, 2025)

Reliabilitas soal Bahasa Arab dari berbagai studi menunjukkan variasi nilai, dengan sebagian besar berada dalam kategori tinggi hingga sangat tinggi. Muhammad Taufiq Ismail & Farikh Marzuki Ammar (2024) menemukan koefisien Cronbach's Alpha 0,861, Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H (2022) sebesar 0,868, Nurul Fikriyah (2021) sebesar 0,854, dan Rahmat Danni,

Ajeng Wahyuni, Tauratiya (2021) sebesar 0,884, yang semuanya masuk dalam kategori sangat tinggi. Sementara itu, Deni Maulana & Anwar Sanusi (2020) menemukan reliabilitas 0,68 (kategori cukup) dan Siti Fathimah Al Fathiyah (2019) menemukan 0,54 (kategori rendah).

Berdasarkan seluruh studi yang dianalisis, satu penelitian menunjukkan reliabilitas rendah, satu reliabilitas cukup, dan empat reliabilitas sangat tinggi. Hal ini mengindikasikan bahwa soal Bahasa Arab yang digunakan di berbagai sekolah umumnya memiliki tingkat reliabilitas yang baik, sehingga dapat dipercaya dalam mengukur kompetensi siswa secara konsisten.

Tingkat Kesukaran

Distribusi soal berdasarkan tingkat kesukaran (mudah, sedang, sukar) dipaparkan dalam tabel berikut.

Tabel 6. Distribusi tingkat kesukaran soal sukar

No.	Author	Tahun	Tingkat Kesukaran	
			sukar	
			Banyak soal	persentase
1	Muhammad Taufiq Ismail & Farikh Marzuki Ammar	2024	0 dari 25 butir soal	0%
3	Siti Fathimah Al Fathiyah	2019	4 dari 15 butir soal	27%
4	Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H	2022	0 dari 50 butir soal	0%
5	Nurul Fikriyah	2021	3 dari 31 butir soal valid	10%
7	Deni Maulana & Anwar Sanusi	2020	3 dari 35 butir soal	9%
8	Rahmat Danni, Ajeng Wahyuni, Tauratiya	2021	0 dari 40 butir soal valid	0%
Rata-rata Persentase				7,67%

(Sumber: Hasil Analisis Data, 2025)

Hasil analisis beberapa penelitian sebagaimana dipaparkan pada tabel, diketahui persentase Tingkat kesukaran soal Bahasa Arab dengan kategori "sulit" yaitu: Muhammad Taufiq Ismail & Farikh Marzuki Ammar (2024) menemukan sebanyak 0 dari 25 (0%) soal yang termasuk kategori sukar, Siti Fathimah Al Fathiyah (2019) menemukan sebanyak 4 dari 15 butir soal (27%) termasuk dalam kategori soal sukar, Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H (2022) menemukan sebanyak 0 dari 50 (0%) soal yang termasuk kategori sukar, Nurul Fikriyah (2021) menemukan 3 dari 31 butir soal valid (10%) termasuk ke dalam kategori sulit, Deni Maulana & Anwar Sanusi (2020) menemukan 3 dari 35 (9%)

butir soal memiliki tingkat kesukaran sukar, Terakhir Rahmat Danni, Ajeng Wahyuni, Tauratiya (2021) menemukan 0 dari 40 butir soal valid (0%) yang termasuk kategori sukar. Jika hasil dari seluruh studi tersebut disimpulkan maka ditemukan bahwa rata-rata soal dengan Tingkat kesukaran "sulit/sukar" sebanyak 7,67%.

Tabel 7. Distribusi tingkat kesukaran soal sedang dan mudah

No	Author	Tahun	Tingkat Kesukaran			
			Sedang		Mudah	
			Banyak soal	persentase	Banyak soal	Persentase
1	Muhammad Taufiq Ismail & Farikh Marzuki Ammar	2024	11 dari 25 butir soal	44%	14 dari 25 butir soal	56%
2	Siti Fathimah Al Fathiyah	2019	3 dari 15 butir soal	20%	8 dari 15 butir soal	53%
3	Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H	2022	7 dari 50 butir soal	14%	43 dari 50 butir soal	86%
4	Nurul Fikriyah	2021	25 dari 31 butir soal valid	80%	3 dari 31 butir soal valid	10%
5	Deni Maulana & Anwar Sanusi	2020	16 dari 35 butir soal	46%	16 dari 35 butir soal	45%
6	Rahmat Danni, Ajeng Wahyuni, Tauratiya	2021	18 dari 40 butir soal valid	45%	22 dari 40 butir soal	55%
Rata-rata Persentase			41,5%.		50,83%	

(Sumber: Hasil Analisis Data, 2025)

Penelitian sebelumnya menunjukkan variasi dalam persentase butir soal Bahasa Arab yang memiliki tingkat kesulitan sedang. Muhammad Taufiq Ismail & Farikh Marzuki Ammar (2024) menemukan 44% soal berada pada kategori ini, sedangkan Siti Fathimah Al Fathiyah (2019) melaporkan 20%. Penelitian Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H (2022) mencatat 14% soal berkategori sedang, sementara Nurul Fikriyah (2021) menemukan 80% butir soal valid memiliki kesulitan sedang. Deni Maulana & Anwar Sanusi (2020) melaporkan 46%, dan Rahmat Danni, Ajeng Wahyuni, & Tauratiya (2021) menemukan 45% soal tergolong sedang. Jika seluruh hasil penelitian ini dirata-ratakan, maka persentase soal dengan tingkat kesulitan sedang adalah 41,5%.

Selain itu, persentase soal Bahasa Arab yang tergolong mudah juga bervariasi. Muhammad Taufiq Ismail & Farikh Marzuki Ammar (2024) menemukan 56% soal termasuk kategori mudah jika digabung dengan kategori sangat mudah, sementara Siti Fathimah Al Fathiyah (2019) melaporkan 53%. Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H (2022) mencatat 86% soal tergolong mudah, sedangkan Nurul Fikriyah (2021) menemukan 10% soal dengan kesulitan mudah. Deni Maulana & Anwar Sanusi (2020) melaporkan 45% soal dalam kategori mudah jika digabung dengan sangat mudah, dan Rahmat Danni, Ajeng Wahyuni, & Tauratiya (2021) menemukan 55% soal tergolong mudah. Rata-rata keseluruhan menunjukkan bahwa 50,83% soal memiliki tingkat kesulitan mudah.

Daya Beda

Kemampuan soal dalam membedakan siswa berkemampuan tinggi dan rendah ditunjukkan pada table berikut.

Tabel 8. Persentase daya beda hasil penelitian

No.	Author	Tahun	Daya Beda			
			Baik		Tidak Baik	
			Banyak soal	Persentase	Banyak soal	Persentase
1	Muhammad Taufiq Ismail & Farikh Marzuki Ammar	2024	tidak menganalisis daya beda		tidak menganalisis daya beda	
2	Siti Fathimah Al Fathiyah	2019	6 dari 15 butir soal	40%	9 dari 15 butir soal	60%
3	Bahrudin Fahmi, Syahrul Rizqi,	2022	26 dari 50 butir soal	52%	24 dari 50 butir soal	48%

	& Nurul Elmira H					
4	Nurul Fikriyah	2021	25 dari 31 butir soal	81%	6 dari 31 butir soal	19%
5	Deni Maulana & Anwar Sanusi	2020	20 dari 35 butir soal	57%	15 dari 35 butir soal	43%
6	Rahmat Danni, Ajeng Wahyuni, Tauratiya	2021	40 dari 40 butir soal	100%	0 dari 40 butir soal	0%
Rata-rata Persentase			66%		34%	

(Sumber: Hasil Analisis Data, 2025)

Penelitian sebelumnya menunjukkan variasi daya beda butir soal Bahasa Arab dalam kategori cukup baik dan baik. Siti Fathimah Al Fathiyah (2019) menemukan 6 dari 15 butir soal (40%) dengan daya beda baik, sementara Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H (2022) melaporkan 26 dari 50 butir soal (52%) dalam kategori yang sama. Nurul Fikriyah (2021) mengidentifikasi 25 dari 31 butir soal (81%) dengan daya beda baik, sedangkan Deni Maulana & Anwar Sanusi (2020) menemukan 20 dari 35 butir soal (57%) dalam kategori baik. Studi oleh Rahmat Danni, Ajeng Wahyuni, & Tauratiya (2021) melaporkan seluruh 40 butir soal valid (100%) memiliki daya beda baik. Jika hasil dari semua penelitian (kecuali Muhammad Taufiq Ismail & Farikh Marzuki Ammar (2024) yang tidak menganalisis daya beda) dirata-ratakan, maka persentase butir soal dengan daya beda baik mencapai 66%.

Selain itu, penelitian juga mengungkap bahwa sebagian butir soal memiliki daya beda kurang baik dan tidak baik. Siti Fathimah Al Fathiyah (2019) menemukan 9 dari 15 butir soal (60%) dalam kategori tidak baik, sementara Bahrudin Fahmi, Syahrul Rizqi, & Nurul Elmira H (2022) mengidentifikasi 24 dari 50 butir soal (48%) dalam kategori rendah hingga sangat rendah. Nurul Fikriyah (2021) melaporkan 6 dari 31 butir soal (19%) memiliki daya beda rendah, sedangkan Deni Maulana & Anwar Sanusi (2020) menemukan 15 dari 35 butir soal (43%) dalam kategori tidak baik. Rahmat Danni, Ajeng Wahyuni, & Tauratiya (2021) tidak menemukan satupun butir soal dalam kategori tidak baik (0%). Jika hasil dari seluruh studi dirata-ratakan (dengan pengecualian studi Muhammad Taufiq Ismail & Farikh Marzuki Ammar (2024)), maka persentase butir soal dengan daya beda tidak baik mencapai 34%.

Temuan Umum

Validitas

Rata-rata 66% butir soal dinyatakan valid, sementara 34% tidak valid. Beberapa tes menunjukkan validitas sangat tinggi (100%), tetapi ada juga yang hanya memiliki validitas sangat rendah (2.85%). Perbedaan yang signifikan ini menunjukkan bahwa tidak semua penyusunan tes Bahasa Arab di Indonesia mengikuti prosedur yang ketat dalam pengujian validitas soal.

Reliabilitas

Sebagian besar penelitian menunjukkan reliabilitas tinggi hingga sangat tinggi dengan nilai ≥ 0.85. Namun, ada penelitian yang menunjukkan reliabilitas rendah (0.54) yang mengindikasikan bahwa hasil tes mungkin kurang konsisten.

Tingkat Kesukaran

Mayoritas butir soal terlalu mudah (50.83%), sementara yang masuk kategori sulit hanya 7.67%. Distribusi tingkat kesukaran yang tidak seimbang ini dapat menyebabkan ketidakakuratan dalam mengukur kompetensi siswa.

Daya Beda

Rata-rata 66% butir soal memiliki daya beda baik, tetapi 34% lainnya memiliki daya beda rendah atau tidak mampu membedakan siswa dengan kemampuan tinggi dan rendah. Jika daya beda rendah, maka soal tidak efektif dalam mengidentifikasi siswa yang benar-benar memahami materi.

Interpretasi Data

Hasil analisis menunjukkan bahwa hanya 66% butir soal yang valid, sedangkan 34% lainnya tidak valid, yang berarti sepertiga soal tidak mampu mengukur kompetensi siswa secara akurat. Hal ini dapat menghasilkan data yang salah atau diinterpretasikan keliru, sehingga berisiko menyebabkan kesalahan dalam pengambilan keputusan pendidikan, intervensi yang keliru, memperburuk ketidaksetaraan pendidikan, dan terhambatnya perkembangan siswa (Mandinach & Schildkamp, 2021). Soal tidak valid juga dapat mengurangi reliabilitas tes secara keseluruhan dan menyumbang pada pemborosan sumber daya karena waktu guru dan siswa terbuang untuk soal yang tidak bermutu (Ramadhan dkk., 2020; Erlinawati & Muslimah, 2021). Penyebab umum dari rendahnya validitas antara lain adalah tidak dilakukannya uji validitas isi atau konstruk sebelum pelaksanaan tes (Mukhlisa, 2023), serta ketidaksesuaian soal dengan aspek linguistik yang seharusnya diukur, khususnya dalam konteks Bahasa Arab (Seelawi dkk., 2021). Oleh karena itu, peningkatan kualitas soal dapat dilakukan melalui validasi awal oleh pakar dan analisis butir pasca-tes guna memastikan bahwa soal yang digunakan benar-benar sesuai dengan tujuan pembelajaran (Karim dkk., 2021).

Data juga menunjukkan bahwa 34% butir soal memiliki daya beda rendah, sehingga gagal membedakan siswa berkemampuan tinggi dan rendah secara akurat. Daya beda yang rendah pada butir soal menyebabkan kegagalan

dalam membedakan siswa dengan tingkat penguasaan materi yang berbeda, sehingga hasil evaluasi menjadi tidak akurat. Kondisi ini dapat menurunkan validitas instrumen asesmen secara keseluruhan dan menghasilkan klasifikasi siswa yang keliru, di mana nilai yang diperoleh tidak lagi mencerminkan tingkat penguasaan materi yang sesungguhnya (Solichin, 2017; Nurhalimah, 2022). Selain itu daya beda yang rendah pada soal asesmen dapat melemahkan efektivitas sistem pembelajaran adaptif karena data yang dihasilkan menjadi bias dan tidak mencerminkan kemampuan siswa secara akurat. Hal ini dapat menyebabkan sistem memberikan rekomendasi pembelajaran yang tidak tepat, sehingga menghambat kemajuan belajar siswa (Kwon dkk., 2023). Penyebab umum rendahnya daya beda soal adalah soal yang terlalu mudah atau terlalu sulit, serta adanya distraktor yang tidak berfungsi optimal sehingga gagal membedakan peserta berkemampuan tinggi dan rendah (Ali & Ruit, 2015; Chauhan et al., 2023; Rezigalla dkk., 2024).

Distribusi tingkat kesukaran soal masih belum seimbang, dengan 50,83% tergolong mudah, 41,5% sedang, dan hanya 7,67% sulit, yang jauh dari distribusi ideal seperti yang disarankan oleh Rezigalla dkk. (2024) yaitu 25% mudah, 50% rata-rata, 25% sulit. Dominasi soal mudah dalam asesmen dapat menyebabkan ilusi kompetensi di kalangan siswa, mengurangi kesiapan mereka untuk menghadapi tantangan yang lebih kompleks, seperti ujian standar nasional atau TOAFL (Talsma, Norris, & Schüz, 2020). Hal ini bisa jadi disebabkan oleh kebijakan pendidikan yang tidak tepat dan desain asesmen yang lemah, yang terlalu fokus pada soal mudah demi mencapai target kelulusan (Panadero dkk., 2019). Akibatnya, asesmen gagal mendeteksi kesenjangan pembelajaran dan siswa kehilangan kesempatan untuk mengembangkan keterampilan metakognitif yang esensial (Phelps, 2012). Oleh karena itu, penting untuk menggunakan Taksonomi Bloom dalam desain soal untuk menyeimbangkan tingkat kesulitan soal, meningkatkan validitas asesmen, dan memastikan soal mampu mengukur kemampuan siswa secara menyeluruh, sehingga siswa lebih siap menghadapi ujian yang lebih kompleks di masa depan (Pizà-Mir, 2022; Abd-Elmoneim dkk., 2023; Koretz, 2024).

SIMPULAN

Hasil penelitian menunjukkan bahwa 66% soal dinyatakan valid dan sebagian besar memiliki reliabilitas tinggi, namun masih ditemukan soal tidak valid dan reliabilitas rendah yang mengindikasikan perlunya penguatan penyusunan soal. Distribusi tingkat kesukaran belum ideal, dengan dominasi soal mudah (50,83%) dan hanya 7,67% soal sulit. Selain itu, 34% soal memiliki daya beda rendah, yang berdampak pada ketidakakuratan asesmen dan kesalahan klasifikasi siswa.

Temuan ini menunjukkan pentingnya validasi awal, pelatihan guru, serta penerapan analisis butir soal secara berkelanjutan untuk memastikan asesmen yang adil, akurat, dan mencerminkan kompetensi siswa secara objektif.

Kontribusi akademik dari penelitian ini terletak pada penyediaan gambaran empiris yang dapat menjadi dasar dalam reformulasi kebijakan evaluasi pembelajaran Bahasa Arab, sekaligus membuka ruang eksplorasi baru dalam pengembangan instrumen penilaian yang lebih adil, berbasis data, dan selaras dengan standar asesmen pendidikan yang terkini serta berorientasi pada peningkatan kualitas pembelajaran.

SARAN DAN REKOMENDASI

Saran untuk Penelitian Selanjutnya

Penelitian berikutnya bisa difokuskan pada analisis soal Bahasa Arab di sekolah, untuk menilai apakah soal sudah sesuai tujuan pembelajaran, bervariasi tingkat kesulitannya, dan mampu membedakan siswa yang menguasai materi dan yang belum. Penelitian juga bisa melihat apakah soal mencakup berbagai tingkat berpikir siswa berdasarkan taksonomi Bloom. Selain itu, penting juga meneliti dampak pelatihan analisis butir soal terhadap peningkatan kualitas soal, karena masih banyak guru yang belum terbiasa melakukan analisis ini.

Rekomendasi Praktis untuk Guru

Guru disarankan secara rutin melakukan analisis butir soal, baik sebelum (uji coba) maupun setelah ujian sumatif/formatif, untuk memastikan kualitas soal. Pelatihan analisis butir—meliputi validitas, reliabilitas, tingkat kesukaran, dan daya beda—perlu diikuti agar guru mampu mengevaluasi soal secara mandiri. Soal sebaiknya disusun dengan variasi tingkat kesulitan mengacu pada taksonomi Bloom, dari pertanyaan dasar hingga analisis mendalam. Hasil analisis soal yang terbukti berkualitas perlu disimpan dalam bank soal untuk digunakan kembali atau dibagikan dengan guru lain, sehingga memperkaya sumber evaluasi yang terstandar. Langkah ini tidak hanya meningkatkan akurasi penilaian tetapi juga efisiensi penyusunan soal di masa depan.

Rekomendasi untuk Pengambil Kebijakan

Kementerian dan instansi terkait dapat mempertimbangkan untuk menerapkan kebijakan analisis butir soal yang disertai dengan program pelatihan berkelanjutan, didukung integrasi software analisis seperti Iteman atau JMetrik guna meningkatkan efisiensi sistem evaluasi sekolah. Badan Standar, Kurikulum, dan Asesmen Pendidikan bersama Lembaga Pendidikan Tenaga Kependidikan dapat mengembangkan bank soal terstandarisasi berdasarkan parameter validitas, reliabilitas, daya beda, dan tingkat kesukaran. Langkah ini diharapkan dapat meningkatkan kualitas asesmen sekaligus mengembangkan kompetensi guru dalam pelaksanaan penilaian berbasis data.

DAFTAR PUSTAKA

Abd-Elmoneim, D. M., Ghandour, H. H., Elrefaie, D. A., & Khodeir, M. S. (2023). Development of an Arabic test for assessment of semantics for the Arabic-speaking children: the Arabic semantic test. The Egyptian Journal of Otolaryngology, 39(1), 49.

Al Fathiyah, S. F. (2019). Analisis butir soal pelajaran Bahasa Arab di MA Roudlotul Ulum Pagak Malang. Tarbiyatuna: Jurnal Pendidikan Ilmiah, 4(1), 76–100.

Ali, S. H., & Ruit, K. G. (2015). The Impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. Perspectives on medical education, 4, 244-251.

Alwinda, R. H. (2020). Pengembangan instrumen berpikir kreatif matematis siswa berdasarkan teori Taksonomi Bloom dan Evans (Skripsi). UIN Syarif Hidayatullah, Jakarta, Indonesia.

Arifianto, M. L., Amin, M. N., Irhamni, A., Ahsanuddin, M., Nikmah, K., Anwar, M. S., & Fitria, N. (2021). Evaluasi pembelajaran dan pengembangan tes interaktif bahasa Arab. Tonggak Media.

Budiono, A. N., & Hatip, M. (2023). Asesmen pembelajaran pada kurikulum merdeka. Jurnal Axioma: Jurnal Matematika Dan Pembelajaran, 8(1), 109-123.

Chauhan, G. R., Chauhan, B. R., Vaza, J. V., & Chauhan, P. R. (2023). Relations of the Number of Functioning Distractors With the Item Difficulty Index and the Item Discrimination Power in the Multiple Choice Questions. Cureus, 15(7), e42492.

Danni, R., Wahyuni, A., & Tauratiya, T. (2021). Item response theory approach: Kalibrasi butir soal penilaian akhir semester mata pelajaran Bahasa Arab. Arabi: Journal of Arabic Studies, 6(1), 93–104.

Erlinawati, E., & Muslimah, M. (2021). Test validity and reliability in learning evaluation. Bulletin of Community Engagement, 1(1), 26-31.

Fahmi, B., Rizqi, S., & Harmeilinda, N. E. (2022). Analisis butir soal Bahasa Arab MAS Pondok Pesantren Assalam Kampar Riau. Ta'lim Al-'Arabiyyah: Jurnal Pendidikan Bahasa Arab & Kebahasaaraban, 6(1), 95–105.

Fatimah, L. U., & Alfath, K. (2019). Analisis kesukaran soal, daya pembeda dan fungsi distraktor. Al-Manar, 8(2), 37–64.

Fikriyah, N. (2021). Analisis butir soal ulangan tengah semester mata pelajaran Bahasa Arab kelas VII semester ganjil SMP Muhammadiyah 1 Yogyakarta tahun ajaran 2019/2020. Maharaat: Jurnal Pendidikan Bahasa Arab, 3(2), 128–140.

Ismail, M. T., & Ammar, F. M. (2024). Analisis butir soal pelajaran Bahasa Arab sumatif akhir semester ganjil tahun ajaran 2022/2023 kelas XI Sekolah

Menengah Atas Al-Fattah Sidoarjo. Jurnal Ilmiah Pendidikan Dasar, 9(2), 3556–3564.

Ismiyati, Raharjo, T. H., Tusyanah, & Sholikah, M. (2023). Pelatihan analisis butir soal berdasarkan teori tes klasik berbantuan Iteman untuk meningkatkan kualitas instrumen penilaian. JAPI, 8(2), 201–210.

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. EduLite: Journal of English Education, Literature and Culture, 6(2), 256-269.

Koretz, D. (2024). "Improving Balance in Educational Measurement: A Legacy of E.F. Lindquist" (Journal of Educational and Behavioral Statistics, 49(6), 930-945)

Kwon, S., Kim, S., Lee, S., Kim, J. Y., An, S., & Kim, K. (2023, October). Addressing Selection Bias in Computerized Adaptive Testing: A User-Wise Aggregate Influence Function Approach. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (pp. 4674-4680).

Mandinach, E. B., & Schildkamp, K. (2021). Misconceptions about data-based decision making in education: An exploration of the literature. Studies in Educational Evaluation, 69, 100842.

Maulana, D., & Sanusi, A. (2020). Analisis butir soal Bahasa Arab Ujian Akhir Madrasah Bersama Daerah (UAMBD) Madrasah Ibtidaiyah tahun 2017–2018. Jurnal Pendidikan Bahasa Arab & Kebahasaaraban, 4(1), 12–24.

Miladya, J. (2021). Evaluasi dalam pembelajaran Bahasa Arab. Dalam Prosiding Konferensi Nasional Bahasa Arab (KONASBARA) (pp. 179–187). Malang, Indonesia.

Mukhlisa, N. (2023). Validitas Tes. JUARA SD: Jurnal Pendidikan Dan Pembelajaran Sekolah Dasar, 2(1), 142-147.

Musa, M. A., Mutiah, R., & Rahmani. (2024). Analisis butir soal Bahasa Arab di MTsN Kota Parepare. Sipakainge, 2(5), 1–10.

Muzaffar, A. (2016). Validitas Tes dan Kualitas Butir Soal. لسـانـنـا (LISANUNA): Jurnal Ilmu Bahasa Arab dan Pembelajarannya, 5(1), 128-143.

Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). Hubungan antara validitas item dengan daya pembeda dan tingkat kesukaran soal pilihan ganda pas. Natural Science Education Research (NSER), 4(3), 249-257.

Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (2019). Using formative assessment to influence self-and co-regulated learning: the role of evaluative judgement. European Journal of Psychology of Education, 34, 535-557.

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. International Journal of Testing, 12(1), 21-43.

Pizà-Mir, B. (2022). Validation of the Use of Bloom's Revised Taxonomy as a Tool for the Design of Assessment Tests. Preprints.

Ramadhan, S., Sumiharsono, R., Mardapi, D., & Prasetyo, Z. K. (2020). The Quality of Test Instruments Constructed by Teachers in Bima Regency, Indonesia: Document Analysis. International Journal of Instruction, 13(2), 507-518.

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., ... & Adam, M. I. E. (2024). Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. BMC Medical Education, 24(1), 445.

Seelawi, H., Tuffaha, I., Gzawi, M., Farhan, W., Talafha, B., Badawi, R., ... & Al-Natsheh, H. (2021, April). ALUE: Arabic language understanding evaluation. In Proceedings of the Sixth Arabic Natural Language Processing Workshop (pp. 173-184).

Sharma, B. (2016). A focus on reliability in developmental research through Cronbach's Alpha among medical, dental and paramedical professionals. Asian Pacific Journal of Health Sciences, 3(4), 271-278.

Solichin, M. (2017). Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan. Dirasat: Jurnal Manajemen Dan Pendidikan Islam, 2(2), 192–213.

Talsma, K., Norris, K., & Schuz, B. (2020). First-year students' academic self-efficacy calibration: Differences by task type, domain specificity, student achievement level, and over time. Student Success, 11(2), 109-121.