AL IBRAH
Journal of Arabic Language Education
Vol. 09, No. 01, June - 2026 | ISSN (P): 2622-4380 | ISSN (E): 2622-6006
https://ejournal.iainpalopo.ac.id/index.php/ibrah

EVALUASI KUALITAS INSTRUMEN TES MAHARAH QIRA'AH MENGGUNAKAN CLASSICAL TEST THEORY

[*1]Sri Wahyuni Hidayanti, [2]Moch Fajarul Falah
[1,2]Universitas Islam Negeri Ar-Raniry
**Corresponding E-mail: 251004004@student.ar-raniry.ac.id*

ARTICLE INFORMATION
Received: 24 May 2026
Revised: 26 May 2026
Accepted: 27 May 2026

DOI:
https://doi.org/10.24256/jale.v9i1.10787

Abstract

This study is motivated by the importance of the quality of assessment instruments in measuring maharah qira'ah (reading comprehension skills) accurately and objectively. The study aims to evaluate the quality of a maharah qira'ah test instrument using the Classical Test Theory (CTT) approach through the analysis of item validity, reliability, difficulty level, and discrimination power. This research employed a descriptive quantitative method, with the object of study being a multiple-choice test instrument consisting of 10 items that had been used in Arabic language learning assessment. The data were obtained from the responses of 17 students from UIN Ar-Raniry and Syiah Kuala University who participated in maharah qira'ah learning at Dayah Darul Aman. Data analysis was conducted using IBM SPSS Statistics 22 through the Corrected Item-Total Correlation (CITC), Cronbach's Alpha, item difficulty index, and item discrimination analysis. The results show that 8 items were valid and 2 items were invalid due to negative CITC values. The reliability coefficient of the instrument was 0.660, which falls into the moderate category. The analysis of item difficulty indicates that 8 items were categorized as easy and 2 items as moderate. Meanwhile, the discrimination index results show that most items were in the good category; however, 2 items had negative discrimination power. The study concludes that the maharah qira'ah test instrument has a fairly good quality, but still requires revision on several items to improve its validity and reliability. The findings indicate that the quality of an assessment instrument is not only determined by the number of valid items, but also by the balance of indicators, the quality of item construction, and the appropriateness of cognitive levels being measured. This study contributes empirical evidence on the importance of item analysis before instruments are used in Arabic language assessment, in order to improve the accuracy of measuring students' reading comprehension skills.

Keywords: *Arabic Learning Evaluation, Classical Test Theory, Maharah Qira'ah, Test Instrument.*

Abstrak

Penelitian ini dilatarbelakangi oleh pentingnya kualitas instrumen evaluasi dalam mengukur kemampuan maharah qira'ah secara akurat dan objektif. Penelitian bertujuan untuk mengevaluasi kualitas instrumen tes maharah qira'ah menggunakan pendekatan Classical Test Theory (CTT) melalui analisis validitas, reliabilitas, tingkat kesukaran, dan daya pembeda butir soal. Penelitian ini menggunakan metode deskriptif kuantitatif dengan objek penelitian berupa instrumen tes pilihan ganda sebanyak 10 butir soal yang telah digunakan dalam evaluasi pembelajaran bahasa Arab. Data penelitian diperoleh dari hasil jawaban 17 mahasiswa UIN Ar-Raniry dan Universitas Syiah Kuala yang mengikuti pembelajaran maharah qira'ah di Dayah Darul Aman. Analisis data dilakukan menggunakan IBM SPSS Statistics 22 melalui uji Corrected Item-Total Correlation (CITC), Cronbach Alpha, indeks kesukaran, dan daya pembeda butir soal. Hasil penelitian menunjukkan bahwa terdapat 8 butir soal valid dan 2 butir soal tidak valid karena memiliki nilai CITC negatif. Nilai reliabilitas instrumen memperoleh Cronbach Alpha sebesar 0,660 yang berada pada kategori cukup. Analisis tingkat kesukaran menunjukkan 8 butir soal berada pada kategori mudah dan 2 butir soal berada pada kategori sedang. Sementara itu, hasil daya pembeda menunjukkan sebagian besar butir soal berkategori baik, namun terdapat 2 butir soal dengan daya pembeda negatif. Penelitian ini menyimpulkan bahwa instrumen tes maharah qira'ah memiliki kualitas yang cukup baik, namun masih memerlukan revisi pada beberapa butir soal agar lebih valid dan reliabel. Temuan penelitian menunjukkan bahwa kualitas instrumen tidak hanya ditentukan oleh jumlah butir soal yang valid, tetapi juga oleh keseimbangan indikator, kualitas konstruksi soal, serta kesesuaian tingkat kognitif yang diukur. Kontribusi penelitian ini memberikan informasi empiris mengenai pentingnya analisis butir soal sebelum instrumen digunakan dalam evaluasi pembelajaran bahasa Arab sehingga dapat meningkatkan kualitas pengukuran kemampuan membaca peserta didik.

Kata Kunci: *Evaluasi Pembelajaran Bahasa Aarab, Classical Test Theory, Maharah Qira'ah,Instrumen Evaluasi.*

PENDAHULUAN

Evaluasi pembelajaran merupakan sebuah proses mengumpulkan, menganalisis, dan menjelaskan informasi tentang hasil belajar siswa, proses pembelajaran, dan program pendidikan secara keseluruhan (Agisna et al., 2023). Tujuan utamanya adalah menilai efektivitas, kualitas, relevansi, dan dampak dari kegiatan pembelajaran atau program pendidikan, serta mengambil keputusan untuk memperbaiki dan meningkatkan pembelajaran sehingga tujuan pendidikan yang telah ditetapkan dapat tercapai (Nazilah & Navlia, 2026). Maka evaluasi pembelajaran merupakan bagian penting dalam proses pendidikan karena berfungsi untuk mengukur ketercapaian tujuan pembelajaran serta mengetahui tingkat kemampuan peserta didik.

Keakuratan hasil evaluasi sangat dipengaruhi oleh kualitas instrumen tes yang digunakan (Putu Gede, 2024). Instrumen yang baik harus memiliki validitas dan reliabilitas agar mampu mengukur kemampuan peserta didik (Armedi, 2025). Instrumen yang tidak valid dan tidak reliabel akan menghasilkan data yang tidak akurat sehingga kesimpulan dan keputusan pendidikan yang diambil dari evaluasi menjadi tidak tepat (Afifah et al., 2025).

Dalam pembelajaran bahasa Arab, maharah qira'ah menjadi maharah yang memiliki evaluasi yang kompleks, karena penilaiannya tidak hanya menuntut kemampuan membaca, tetapi juga mencakup pemahaman teks yang komprehensif. Penilaian tersebut meliputi memahami kosa kata, menemukan informasi utama, menafsirkan isi bacaan, serta memahami antar gagasan dalam teks bahasa Arab (Lazuardi et al., 2025). Kompleksitas ini menunjukkan bahwa dalam mengevaluasi pembelajaran maharah qira'ah diperlukan instrumen tes yang berkualitas agar hasil pengukuran kemampuan membaca siswa dapat dilakukan secara akurat.

Meskipun instrumen evaluasi memegang peranan penting dalam pembelajaran, pengujian kualitas butir soal maharah qira'ah dalam praktik evaluasi pembelajaran masih belum dilakukan secara optimal. Guru seringkali langsung menggunakan soal evaluasi tanpa menguji validitas dan reliabilitas terlebih dahulu sehingga kualitas instrumen belum diketahui secara pasti (Lazuardi et al., 2025). Kondisi ini berpotensi menyebabkan butir soal yang digunakan mengukur aspek yang tidak relevan dengan indikator pembelajaran, dan terjadinya kesalahan pengambilan keputusan hasil evaluasi pembelajaran (Handoko et al., 2025). Untuk mengetahui kualitas instrumen yang digunakan, diperlukan analisis secara sistematis terhadap validitas, reliabilitas, tingkat kesukaran, dan daya pembeda butir soal pada tes maharah qira'ah.

Classical Test Theory (CTT) menjadi salah satu pendekatan dalam psikometri yang dapat digunakan untuk menganalisis kualitas tes. Teori ini menjelaskan bahwa skor yang diperoleh peserta didik merupakan hasil dari skor kemampuan sebenarnya yang dipengaruhi oleh kesalahan pengukuran (Sumaryanta, 2021). Dalam teori ini, kualitas instrumen dipengaruhi oleh sejauh mana butir soal mampu merepresentasikan kemampuan yang diukur secara konsisten dengan tingkat kesalahan pengukuran yang minimal (Suseno & Susongko, 2021). Oleh karena itu, analisis kualitas instrumen dalam pendekatan *Classical Test Theory (CTT)* dilakukan melalui pengujian validitas, reliabilitas, tingkat kesukaran, dan daya pembeda butir soal. Melalui pendekatan ini, kelayakan instrumen dalam mengukur kemampuan maharah qira'ah dapat diketahui secara lebih objektif.

Penelitian mengenai analisis instrumen tes maharah qira'ah dengan pendekatan *Classical Test Theory (CTT)* telah dilakukan dalam berbagai konteks pembelajaran bahasa Arab. Beberapa penelitian sebelumnya membahas pengembangan tes berbasis media pembelajaran(Damogalad et al., 2024), analisis kualitas butir soal pada Madrasah Aliyah(Arief Maulana, 2025), serta pengujian validitas dan reliabilitas instrumen evaluasi tingkat pendidikan dasar(Hilmi et al., 2025). Namun, sebagian besar penelitian tersebut menganalisis instrumen yang telah melalui tahap pengembangan dan validasi sebelum digunakan dalam evaluasi pembelajaran.

Berbeda dengan penelitian sebelumnya, penelitian ini memfokuskan analisis pada instrumen tes maharah qira'ah yang telah digunakan secara langsung dalam evaluasi pembelajaran tanpa melalui pengujian empiris terlebih dahulu. Padahal, instrumen yang tidak diuji kualitas butirnya secara empiris berpotensi menimbulkan bias pengukuran sehingga belum tentu mampu merepresentasikan indikator kemampuan maharah qira'ah secara utuh (Arief Maulana, 2025). Selain itu, penelitian ini dilakukan pada konteks pembelajaran bahasa Arab tingkat mahasiswa di lingkungan dayah yang masih jarang dikaji dalam penelitian analisis instrumen berbasis *Classical Test Theory (CTT)*. Oleh karena itu, penelitian ini tidak hanya menilai kualitas teknis instrumen, tetapi juga memberikan gambaran empiris mengenai kelayakan instrumen evaluasi yang digunakan dalam praktik pembelajaran bahasa Arab di lingkungan dayah.

Berdasarkan latar belakang tersebut, penelitian ini dilakukan untuk menjawab pertanyaan mengenai bagaimana kualitas instrumen tes maharah qira'ah berdasarkan analisis validitas, reliabilitas, tingkat kesukaran, dan daya pembeda butir soal menggunakan pendekatan *Classical Test Theory (CTT).* Penelitian ini bertujuan untuk mengevaluasi kualitas instrumen tes maharah qira'ah berdasarkan pendekatan *Classical Test Theory (CTT)*. Hasil penelitian ini diharapkan dapat memberikan informasi empiris mengenai kualitas instrumen tes maharah qira'ah sehingga dapat digunakan sebagai dasar perbaikan dan pengembangan butir soal evaluasi pembelajaran bahasa Arab, khususnya pada maharah qira'ah yang lebih akurat, objektif, dan sesuai dengan tujuan pembelajaran.

METODE

Jenis penelitian ini adalah penelitian deskriptif kuantitatif, yaitu penelitian yang menggambarkan fakta, kejadian, atau kondisi yang terjadi tanpa adanya manipulasi variabel penelitian.(Waruwu et al., 2025) Objek penelitian berupa instrumen tes maharah qira'ah berbentuk pilihan ganda sebanyak 10 butir soal yang telah digunakan dalam evaluasi pembelajaran bahasa Arab. Data penelitian diperoleh dari hasil jawaban 17 mahasiswa UIN Ar-Raniry dan Universitas Syiah Kuala yang mengikuti program pembelajaran dan evaluasi maharah qira'ah di Dayah Darul Aman dan telah menggunakan instrumen tersebut.

Dayah Darul Aman dipilih sebagai lokasi penelitian karena merupakan lembaga pendidikan berbasis pesantren yang menerapkan pembelajaran bahasa Arab secara intensif melalui pembiasaan membaca dan memahami teks berbahasa Arab dalam kegiatan pembelajaran sehari-hari. Konteks ini relevan karena kemampuan qira'ah di lingkungan dayah tidak hanya menekankan kelancaran membaca, tetapi juga pemahaman makna teks secara langsung.

Adapun instrumen tes yang dianalisis mencakup empat indikator kemampuan maharah qira'ah, yaitu memahami kosa kata, menemukan informasi utama, menafsirkan isi bacaan, serta memahami hubungan antar gagasan dalam teks bahasa Arab mengenai الإحتفال بمولد الرسول seperti yang terdapat dalam table berikut :

Tabel 1. Distribusi Butir Soal Kemampuan Qira'ah

Indikator	Taburan Item	Bilangan
Memahami kosa kata	Soal no.6	1
Menemukan informasi utama	Soal no.1, Soal no.5, Soal no.10	3
Menafsirkan isi bacaan	Soal no.2, Soal no.3, Soal no.8	3
Memahami antar gagasan dalam teks	Soal no.4, Soal no.7, Soal no.9	3
Jumlah		**10**

Teknik pengumpulan data dilakukan melalui dokumentasi instrumen tes maharah qira'ah yang telah digunakan dalam evaluasi pembelajaran beserta hasil jawaban responden terhadap setiap butir soal. Data penelitian berupa skor jawaban responden dengan kategori benar diberi skor 1 dan salah diberi skor 0. Data kemudian diolah menggunakan program IBM SPSS Statistics 22.

Teknik analisis data menggunakan pendekatan *Classical Test Theory (CTT)*. Pendekatan ini dipilih karena melihat skor tes sebagai gabungan antara skor sebenarnya (true score) dan kesalahan pengukuran, serta menilai kualitas butir soal berdasarkan hubungan antara skor setiap butir dengan skor total. Sementara itu, Item Response Theory (IRT) bekerja dengan cara yang lebih kompleks, yaitu memperkirakan kemampuan peserta dan karakteristik setiap butir soal secara terpisah menggunakan model probabilitas, sehingga membutuhkan jumlah responden yang besar dan perhitungan yang lebih rumit agar hasilnya stabil (Purwa Antara, 2020).

Dalam penelitian ini, CTT lebih sesuai digunakan karena tujuan utama adalah mengevaluasi kualitas butir soal secara deskriptif melalui analisis validitas, reliabilitas, tingkat kesukaran, dan daya pembeda, bukan memodelkan kemampuan peserta secara mendalam seperti pada IRT. Selain itu, jumlah responden yang terbatas (17 orang) membuat CTT lebih tepat karena hasilnya masih dapat dipercaya dalam kondisi sampel kecil, sedangkan IRT cenderung tidak stabil jika data sedikit. Penggunaan sampel kecil dalam penelitian berbasis CTT juga dapat dibenarkan secara metodologis apabila konteks penelitian bersifat terbatas dan tertutup, seperti pada kelompok belajar di lingkungan dayah yang tidak memungkinkan penambahan responden di luar populasi yang tersedia. Karakaya dan Alparslan (2022) menjelaskan bahwa dalam studi reliabilitas berbasis Cronbach Alpha, ukuran sampel yang kecil masih dapat memberikan estimasi yang bermakna apabila interpretasinya dilakukan secara hati-hati dan disertai pengakuan atas keterbatasan tersebut. Dengan demikian, 17 responden dalam penelitian ini merupakan keseluruhan populasi yang mengikuti evaluasi maharah qira'ah di Dayah Darul Aman pada periode pengambilan data, sehingga penelitian ini bersifat sensus populasi terbatas, bukan pengambilan sampel acak.

Kemudian pada validitas dan daya pembeda butir dianalisis menggunakan *Corrected Item-Total Correlation (CITC)*, reliabilitas menggunakan koefisien Cronbach Alpha, dan tingkat kesukaran menggunakan indeks kesukaran berdasarkan proporsi jawaban benar. Hasil analisis selanjutnya dikategorikan untuk menentukan kualitas dan kelayakan instrumen dalam mengukur kemampuan maharah qira'ah.

HASIL DAN DISKUSI

Uji Validitas

Validitas instrumen adalah tingkat ketepatan dan kesahihan alat ukur dalam mengungkap data sesuai fakta atau konstruk yang diukur.(Djaali & Mulyono, 2008) Uji Validitas dilakukan untuk mengetahui tingkat ketepatan setiap butir soal dalam mengukur kemampuan pada evaluasi pembelajaran. Pada penelitian ini validitas instrumen evaluasi pembelajaran maharah qira'ah dianalisis melalui *Corrected Item-Total Correlation*

(CITC). Butir soal dinyatakan valid apabila memperoleh nilai *Corrected Item-Total Correlation (CITC)* ≥ 0,30 (Hendriyani, 2021).

Tabel 2. Hasil Uji Validitas Instrumen

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach Alpha if Item Deleted
Soal_1	7.12	3.235	.473	.606
Soal_2	7.06	3.309	.529	.603
Soal_3	7.06	3.434	.418	.622
Soal_4	7.18	3.279	.371	.626
Soal_5	7.29	2.971	.498	.593
Soal_6	7.12	4.110	-.129	.716
Soal_7	7.18	4.029	-.092	.719
Soal_8	7.18	3.029	.551	.584
Soal_9	7.00	3.625	.406	.632
Soal_10	7.29	3.096	.416	.615

Hasil analisis Corrected Item-Total Correlation (CITC) menunjukkan bahwa setiap soal memiliki tingkat kontribusi yang berbeda dalam mengukur kemampuan maharah qira'ah. Secara umum, sebagian besar soal memiliki nilai korelasi positif di atas batas 0,30, yang berarti soal-soal tersebut sudah sesuai dan cukup baik dalam mengukur kemampuan siswa secara konsisten. Selain itu, terdapat perbedaan kekuatan antarsoal, sehingga tidak semua soal memiliki pengaruh yang sama terhadap hasil tes. Di sisi lain, terdapat dua soal yang memiliki nilai korelasi negatif (-0,129 dan -0,092), yang menunjukkan bahwa kedua soal tersebut tidak sejalan dengan keseluruhan tes dan dapat mengurangi kualitas instrumen. Dengan demikian, dapat disimpulkan bahwa tidak semua soal memiliki kualitas yang sama, sehingga beberapa soal perlu diperbaiki agar instrumen menjadi lebih baik.

Uji Reliabilitas

Reliabilitas adalah konsistensi instrumen dalam menghasilkan skor yang relatif sama apabila digunakan pada kondisi atau subjek yang sama pada waktu berbeda atau dengan penilai berbeda, sehingga instrumen dapat dipercaya sebagai alat pengumpul data yang stabil dan dapat diulang.(Djaali & Mulyono, 2008) Uji reliabilitas dilakukan untuk mengetahui tingkat konsistensi dan kestabilan instrumen penelitian. Analisis reliabilitas menggunakan pendekatan internal consistency melalui koefisien *Cronbach Alpha* mencapai 0,70, dengan kriteria sebagai berikut(Kilic, 2016) :

Tabel 3. Kriteria Tingkat Reliabilitas

Cronbach Alpha	**Kategori**
< 0,60	Rendah
0,60–0,69	Cukup
≥ 0,70	Baik/Dapat Diterima
≥ 0,80	Sangat Baik
≥ 0,90	Sangat Tinggi

Tabel 4. Hasil Uji Reliabilitas Instrumen

Reliability Statistics

Cronbach Alpha	N of Items
.660	10

Berdasarkan hasil uji diperoleh nilai sebesar 0,660 pada 10 butir soal. Nilai tersebut menunjukkan bahwa instrumen memiliki tingkat reliabilitas yang cukup, meskipun belum mencapai kategori reliable tinggi dengan standar ≥ 0,70.

Uji tingkat kesukaran

Uji tingkat kesukaran merupakan analisis yang dilakukan untuk mengetahui tingkat kemudahan atau kesulitan suatu butir soal. Analisis ini bertujuan untuk menentukan kategori butir soal dari yang mudah, sedang, atau sukar sehingga dapat diketahui kualitas butir soal dalam instrumen evaluasi pembelajaran. Nilai mean pada setiap butir soal digunakan sebagai indeks kesukaran. Mean yang tinggi menunjukkan bahwa sebagian besar peserta didik menjawab benar sehingga soal tergolong mudah, sedangkan mean yang rendah menunjukkan bahwa soal tergolong sukar(Lastrijanah et al., 2017), dengan kriteria sebagai berikut(Arikunto, 2018) :

Tabel 5. Kriteria Tingkat Kesukaran

Indeks Kesukaran	Kategori
≤0,30	Sukar
0,31–0,70	Sedang
≥0,71	Mudah

Tabel 6. Hasil Uji Tingkat Kesukaran

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Soal_1	17	0	1	.82	.393
Soal_2	17	0	1	.88	.332
Soal_3	17	0	1	.88	.332
Soal_4	17	0	1	.76	.437
Soal_5	17	0	1	.65	.493
Soal_6	17	0	1	.82	.393
Soal_7	17	0	1	.76	.437
Soal_8	17	0	1	.76	.437
Soal_9	17	0	1	.94	.243
Soal_10	17	0	1	.65	.493
Valid N (listwise)	17				

Berdasarkan hasil uji tingkat kesukaran, diperoleh bahwa sebagian besar butir soal berada pada kategori mudah. Dari 10 butir soal yang dianalisis, terdapat 8 butir soal berkategori mudah, yaitu soal nomor 1,2,3,4,6,7,8, dan 9 dengan nilai indeks kesukaran berkisar antara 0,76-0,94. Sementara itu, soal nomor 5 dan 10 berada pada kategori sedang dengan nilai indeks kesukaran sebesar 0,65. Tidak terdapat butir soal yang termasuk kategori sukar. Nilai indeks kesukaran tertinggi terdapat pada soal nomor 9 sebesar 0,94, sedangkan nilai terendah terdapat pada soal nomor 5 dan 10 sebesar 0,65. Hasil tersebut menunjukkan bahwa sebagian besar responden mampu menjawab butir soal dengan benar sehingga instrumen cenderung memiliki tingkat kesukaran yang rendah.

Uji Daya Pembeda Butir Soal

Uji daya pembeda butir soal merupakan analisis yang digunakan untuk mengetahui kemampuan suatu butir soal dalam membedakan peserta didik yang memiliki kemampuan tinggi dan rendah.(Zaenal, 2017) Analisis ini bertujuan untuk mengetahui apakah setiap butir soal mampu membedakan peserta didik yang memahami materi dengan baik dan peserta didik yang kurang memahami materi pada pembelajaran. Adapun kriterianya sebagai berikut(Arikunto, 2018):

Tabel 7. Kriteria Daya Pembeda Butir Soal

Daya Pembeda	Kategori
≥0,40	Baik
0,30–0,39	Cukup
0,20–0,29	Kurang
< 0,20	Tidak Memadai

Tabel 8. Hasil Uji Daya Pembeda Butir Soal

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach Alpha if Item Deleted
Soal_1	7.12	3.235	.473	.606
Soal_2	7.06	3.309	.529	.603
Soal_3	7.06	3.434	.418	.622
Soal_4	7.18	3.279	.371	.626
Soal_5	7.29	2.971	.498	.593
Soal_6	7.12	4.110	-.129	.716
Soal_7	7.18	4.029	-.092	.719
Soal_8	7.18	3.029	.551	.584
Soal_9	7.00	3.625	.406	.632
Soal_10	7.29	3.096	.416	.615

Perlu dijelaskan bahwa dalam pendekatan *Classical Test Theory (CTT)*, nilai *Corrected Item-Total Correlation (CITC)* berfungsi ganda: sebagai indikator validitas butir soal sekaligus sebagai indikator daya pembeda. Hal ini merupakan karakteristik metodologis CTT, di mana korelasi antara skor butir dengan skor total mencerminkan sejauh mana butir soal mengukur konstruk yang sama (validitas internal) sekaligus kemampuannya membedakan peserta berkemampuan tinggi dan rendah (daya pembeda). Oleh karena itu, data yang ditampilkan pada Tabel 8 berikut merupakan data CITC yang sama dengan Tabel 2, namun interpretasinya diarahkan pada aspek daya pembeda sesuai dengan kriteria pada Tabel 7. Penggunaan satu tabel output untuk dua tujuan interpretasi ini adalah lazim dalam penelitian CTT dan tidak mengurangi validitas analisis.

Berdasarkan hasil analisis *Corrected Item-Total Correlation (CITC)* dalam pendekatan *Classical Test Theory (CTT)* yang digunakan sebagai indikator daya pembeda butir soal, diperoleh bahwa sebagian besar butir soal memiliki daya pembeda kategori baik. Sebanyak 7 butir soal, yaitu soal 1, 2, 3, 5, 8, 9, dan 10, berada pada kategori baik dengan nilai korelasi di atas 0,40, sedangkan soal 4 berada pada kategori cukup dengan nilai 0,371. Namun terdapat 2 butir soal yang memiliki daya pembeda sangat rendah, yaitu soal 6 dan 7 dengan korelasi negatif masing-masing -0,129 dan -0,092.

Hasil uji validitas diperoleh bahwa dari 10 butir soal terdapat 8 butir soal yang dinyatakan valid, yaitu soal nomor 1, 2, 3, 4, 5, 8, 9, dan 10 dengan rentang nilai 0,371–0,551. Nilai tersebut menunjukkan bahwa butir-butir soal memiliki keterkaitan yang cukup baik dengan skor total sehingga mampu mengukur konstruk maharah qira'ah secara konsisten. Sementara itu, 2 butir soal lainnya, yaitu nomor 6 dan 7, dinyatakan tidak valid karena memiliki nilai korelasi negatif masing-masing sebesar -0,129 dan -0,092. Dalam teori konsistensi internal yang dikembangkan oleh Lee J. Cronbach, korelasi negatif menunjukkan bahwa suatu butir tidak sejalan dengan keseluruhan tes dan berpotensi mengukur kemampuan yang berbeda, sehingga perlu direvisi atau dieliminasi agar kualitas instrumen menjadi lebih baik (Keith, 2018).

Secara indikator, soal nomor 6 yang mengukur kemampuan memahami kosa kata menunjukkan nilai korelasi negatif meskipun memiliki tingkat kesukaran yang rendah. Dalam teori pemerolehan kosakata, kemampuan kosakata tidak hanya berkaitan dengan pengenalan bentuk kata, tetapi juga pemahaman makna dalam konteks (Hashim et al., 2020). Tingkat kesukaran yang terlalu rendah menunjukkan bahwa soal cenderung terlalu mudah sehingga peserta didik dengan kemampuan tinggi maupun rendah dapat menjawab dengan pola yang hampir sama. Kondisi tersebut menyebabkan butir soal tidak mampu membedakan kemampuan peserta didik secara optimal dan menghasilkan korelasi negatif terhadap skor total. Selain itu, indikator memahami kosa

kata hanya diwakili oleh satu butir soal sehingga kesalahan pada satu item memberikan pengaruh besar terhadap kestabilan pengukuran indikator tersebut (Zhang & Colvin, 2024).

Hal serupa juga ditemukan pada soal nomor 7 yang mengukur kemampuan memahami hubungan antar gagasan dalam teks. Meskipun kemampuan memahami hubungan antar gagasan merupakan keterampilan membaca tingkat tinggi karena melibatkan proses menghubungkan informasi antarkalimat(Ambarita et al., 2021), soal ini justru berada pada kategori mudah. Akibatnya, soal tidak cukup menantang dan kurang mampu membedakan peserta didik yang memiliki kemampuan tinggi dan rendah. Kondisi tersebut menyebabkan jawaban responden menjadi kurang konsisten dengan konstruk yang diukur sehingga nilai korelasi butir menjadi negatif.

Pada uji reliabilitas, diperoleh nilai *Cronbach Alpha* sebesar 0,660 yang menunjukkan bahwa instrumen memiliki tingkat konsistensi internal pada kategori cukup. Walaupun belum mencapai standard ideal ≥ 0,70, nilai ini masih menunjukkan bahwa instrumen relative konsisten dalam mengukur kemampuan maharah qira'ah. Hasil ini sejalan dengan penelitian Citra dkk yang menunjukkan bahwa beberapa kelompok instrumen memperoleh nilai *Cronbach Alpha* pada kategori cukup masih dapat digunakan dalam pengembangan instrumen evaluasi (Cesilia et al., 2021). Nilai reliabilitas yang belum optimal dapat dipengaruhi oleh jumlah butir soal yang terbatas, jumlah responden yang relatif kecil, serta keberadaan butir soal tidak valid yang turut menurunkan konsistensi instrumen secara keseluruhan. Perlu dicatat bahwa instrumen yang dianalisis dalam penelitian ini hanya terdiri dari 10 butir soal, yang merupakan keterbatasan inheren dari instrumen yang sedang dievaluasi, bukan dari desain penelitian. Secara psikometrik, instrumen dengan butir soal yang lebih sedikit cenderung menghasilkan estimasi reliabilitas yang lebih rendah karena cakupan konten yang lebih sempit. Hal ini menjadi salah satu temuan penting penelitian ini, yakni bahwa instrumen maharah qira'ah yang hanya terdiri dari 10 butir soal perlu dikembangkan lebih lanjut dengan menambah jumlah item agar reliabilitas dan representasi indikator kemampuan membaca dapat meningkat. (Karakaya & Alparslan, 2022) Oleh karena itu, revisi terhadap butir soal yang lemah diperlukan agar reliabilitas instrumen dapat meningkat pada pengujian berikutnya.

Selanjutnya, hasil analisis tingkat kesukaran menunjukkan bahwa sebagian besar butir soal berada pada kategori mudah, yaitu 8 butir soal dengan indeks kesukaran 0,76–0,94, sedangkan 2 butir lainnya berada pada kategori sedang dengan nilai 0,65. Tidak terdapat butir soal yang termasuk kategori sukar. Jika dibandingkan dengan standar ideal penyusunan tes menurut Arikunto, komposisi tingkat kesukaran yang baik seharusnya terdiri dari sekitar 25% soal sukar, 50% sedang, dan 25% mudah (Arikunto, 2018). Dengan demikian, distribusi tingkat kesukaran dalam instrumen maharah qira'ah ini belum memenuhi standar tersebut karena didominasi oleh soal mudah dan tidak adanya soal pada kategori sukar. Dominasi butir soal mudah tersebut menunjukkan bahwa instrumen belum memberikan variasi tingkat kesulitan yang optimal, sehingga kemampuan tes dalam membedakan peserta didik berkemampuan tinggi dan rendah menjadi terbatas. Sebagaimana dijelaskan oleh Assad dkk, komposisi soal yang terlalu mudah dapat mengurangi kemampuan instrumen dalam membedakan kemampuan responden secara lebih akurat (Rezigalla et al., 2024).

Dalam pendekatan *Classical Test Theory (CTT)*, tingkat kesukaran dan daya pembeda memiliki hubungan yang erat karena butir soal yang terlalu mudah cenderung memiliki kemampuan diskriminasi yang rendah (Purwa Antara, 2020). Kondisi tersebut terlihat pada soal nomor 6 dan 7 yang memiliki indeks kesukaran tinggi, tetapi menunjukkan korelasi item-total dan daya pembeda bernilai negatif. Hal ini menunjukkan bahwa kedua soal tidak mampu membedakan peserta didik berkemampuan tinggi dan rendah, bahkan berpotensi berlawanan dengan tujuan pengukuran.(Yustiandi & Saepuzaman, 2024) Penelitian lain yang dilakukan oleh Sri Rejeki dkk juga menjelaskan bahwa item dengan discrimination index buruk perlu direvisi karena tidak efektif sebagai alat evaluasi pembelajaran (Rejeki et al., 2023).

Secara keseluruhan, instrumen tes maharah qira'ah ini tergolong cukup baik, tetapi masih memiliki beberapa kelemahan pada tingkat berpikir soal, tingkat kesukaran, dan adanya beberapa soal yang memiliki korelasi negatif. Jika dilihat dari Taksonomi Bloom, sebagian besar soal berada pada level C1 (mengingat) dan C2 (memahami), seperti soal nomor 6 tentang kosakata serta soal nomor 1, 2, 3, 4, 5, 7, 8, 9, dan 10 yang hanya meminta pemahaman isi teks secara langsung. Hal ini menunjukkan bahwa soal masih didominasi kemampuan berpikir tingkat rendah, sehingga belum sepenuhnya mengukur kemampuan maharah qira'ah pada tingkat yang lebih tinggi seperti C3 (menerapkan) dan C4 (menganalisis). Kategorisasi level kognitif setiap butir soal berdasarkan Taksonomi Bloom ditampilkan pada Tabel 9 berikut.

Tabel 9. Kategorisasi Level Kognitif Butir Soal Berdasarkan Taksonomi Bloom

No. Soal	Indikator	Level Bloom	Kode	Kategori
1	Menemukan informasi utama	Memahami (Understanding)	C2	LOTS
2	Menafsirkan isi bacaan	Memahami (Understanding)	C2	LOTS
3	Menafsirkan isi bacaan	Memahami (Understanding)	C2	LOTS

4	Memahami antar gagasan	Memahami (Understanding)	C2	LOTS
5	Menemukan informasi utama	Memahami (Understanding)	C2	LOTS
6	Memahami kosa kata	Mengingat (Remembering)	C1	LOTS
7	Memahami antar gagasan	Memahami (Understanding)	C2	LOTS
8	Menafsirkan isi bacaan	Memahami (Understanding)	C2	LOTS
9	Memahami antar gagasan	Memahami (Understanding)	C2	LOTS
10	Menemukan informasi utama	Memahami (Understanding)	C2	LOTS

Hasil ini juga sejalan dengan temuan tingkat kesukaran yang menunjukkan bahwa sebagian besar soal tergolong mudah. Kondisi ini terjadi karena soal lebih banyak menuntut pemahaman langsung terhadap isi teks dan informasi yang sudah jelas, sehingga siswa bisa menjawab dengan cukup melihat kata kunci tanpa perlu berpikir mendalam. Selain itu, pilihan jawaban yang cukup jelas juga membuat soal lebih mudah dikerjakan. Dengan demikian, rendahnya tingkat kesukaran tidak hanya disebabkan oleh bentuk soal, tetapi juga oleh cara pilihan jawaban disusun. Oleh karena itu, instrumen ini masih perlu diperbaiki dengan menambah soal pada level C3–C4 dan menyeimbangkan tingkat kesukaran agar mampu mengukur kemampuan maharah qira'ah dengan lebih baik dan menyeluruh.

KESIMPULAN

instrumen tes maharah qira'ah yang dianalisis menggunakan pendekatan *Classical Test Theory* (CTT) menunjukkan kualitas yang cukup baik, meskipun masih terdapat beberapa kelemahan yang perlu diperbaiki. Hasil uji validitas menunjukkan bahwa dari 10 butir soal terdapat 8 butir soal yang valid dan 2 butir soal tidak valid, yaitu soal nomor 6 dan 7 karena memiliki nilai *Corrected Item-Total Correlation (CITC)* negatif. Pada uji reliabilitas diperoleh nilai *Cronbach Alpha* sebesar 0,660 yang menunjukkan bahwa instrumen memiliki tingkat konsistensi internal pada kategori cukup dan masih dapat digunakan dalam tahap pengembangan instrumen. Hasil analisis tingkat kesukaran menunjukkan bahwa sebagian besar butir soal berada pada kategori mudah sehingga instrumen cenderung kurang mampu memberikan variasi tingkat kemampuan peserta didik secara optimal. Sementara itu, hasil uji daya pembeda menunjukkan bahwa sebagian besar soal memiliki daya pembeda baik, namun terdapat 2 butir soal dengan daya pembeda negatif yang menunjukkan bahwa soal tersebut tidak mampu membedakan peserta didik berkemampuan tinggi dan rendah. Dengan demikian, instrumen evaluasi maharah qira'ah ini telah memenuhi sebagian besar kriteria kualitas instrumen, tetapi masih memerlukan revisi pada butir-butir yang bermasalah agar instrumen menjadi lebih valid, reliabel, dan sensitif dalam mengukur kemampuan membaca bahasa Arab.

Berdasarkan temuan penelitian, pengembangan instrumen evaluasi maharah qira'ah disarankan menggunakan jumlah sampel uji coba yang lebih memadai, minimal 30–40 responden, agar hasil analisis validitas dan reliabilitas lebih stabil. Selain itu, distribusi tingkat kesukaran perlu dibuat lebih proporsional dengan menyediakan variasi soal mudah, sedang, dan sukar sehingga instrumen mampu membedakan kemampuan peserta didik secara lebih optimal. Setiap indikator kemampuan membaca juga perlu diwakili oleh minimal 2–3 butir soal agar pengukuran lebih konsisten dan representatif. Butir soal yang memiliki nilai validitas dan daya pembeda negatif perlu direvisi atau dieliminasi sebelum instrumen digunakan dalam evaluasi pembelajaran. Penelitian selanjutnya juga disarankan analisis instrumen dengan pendekatan psikometri lain seperti *Item Response Theory* (IRT) agar diperoleh hasil analisis yang lebih mendalam dan akurat terhadap kualitas instrumen maharah qira'ah.

DAFTAR PUSTAKA

Afifah, Aulia, Hayatun Zayrin, Khalista Nupus, Maizia Khansa, Marsela Siska, Hidayatullah Rully, and Harmonedi. "Analisis Instrumen Penelitian Pendidikan (Uji Validitas Dan Relibilitas Instrumen Penelitian)." *Jurnal QOSIM Jurnal Pendidikan Sosial & Humaniora* 3, no. 2 (2025).

Agisna, Robi, Zulfikri Alwi Jauhari, M. Saifudin Zuar, Muhammad Sholihin, and Anis Khusnul I. "Evaluasi Pembelajaran." *Social Science Academic* 1, no. 2 (2023): 353–62. https://doi.org/10.37680/ssa.v1i2.3582.

Agung Purwa Antara, Anak. *Penyetaraan Vertikal Dengan Pendekatan Klasik Dan Item Response Theory*. Edited by Avinda Yuda Wati. 1st ed. Yogyakarta: CV Budi Utama, 2020.

Ambarita, Rahel Sonia, Neneng Sri Wulan, and D Wahyudin. "Analisis Kemampuan Membaca Pemahaman Pada Siswa Sekolah Dasar." *Edukatif : Jurnal Ilmu Pendidikan* 3, no. 5 (2021): 2336–44.

https://doi.org/10.31004/edukatif.v3i5.836.

Arief Maulana, Ikhsan. " تحليل بنود أسئلة الاختبار التحصيلي لمهارة القراءة بمعيار ،روبرت إبل' في مدرسة يافينا الثانوية الإسلامية بمدينة لهوكسيوماوي بأتشيه." جامعة موالان مالك إبراهيم الإسالمية احلكومية, 2025.

Armedi, Rama. "Karakteristik Tes Yang Baik Dan Proses Penyusunan Instrumen Tes Untuk Pembelajaran Di Sekolah." *PENDAGOGIA: Jurnal Pendidikan Dasar* 5, no. April (2025): 10–17.

Cesilia, Citra, Diah Asri Wulandari, and Meita Dhamayanti. "Validitas Dan Reliabilitas Ages & Stages Questionnaire: Social-Emotional 2 Versi Indonesia Validitas Dan Reliabilitas Ages & Stages Questionnaire: Social-Emotional 2 Versi Indonesia." *Sari Pediatri* 22, no. 6 (2021): 343. https://doi.org/10.14238/sp22.6.2021.343-50.

Damogalad, Nurul Aulia, Ibnu Rawandhy N Hula, and Nurul Aini Pakaya. "Design and Development of Maharah Qira'ah Test Using the Kahoot Application." *Jurnal Naskhi Jurnal Kajian Pendidikan Dan Bahasa Arab* 6, no. 2 (2024): 38–55. https://doi.org/10.47435/naskhi.v6i2.3062.

Djaali, and Pudji Mulyono. *Pengukuran Dalam Bidang Pendidikan*. *Grasindo Publisher*. 1st ed. Jakarta: PT Gramedia Widiasarana, 2008.

Handoko, Fatkhur Rohman, Syahla Athiya Farha, Hana Salsabila Putri, and Diah Ayu Sucitra. "Optimalisasi Kualitas Butir Soal Melalui Pelatihan Penyusunan Indikator Dengan Uji Validitas Dan Reliabilitas." *Jurnal Pengabdian Masyarakat Ilmu Pendidikan* 4, no. 2 (2025): 296–305. https://doi.org/https://doi.org/10.23960/jpm-ip.vol.4i.2.1083.

Hashim, Hasnurol, Kaseh Abu Bakar, and Maheram Ahmad. "Penguasaan Kosa Kata Bahasa Arab Menerusi Pengetahuan Makna Dan Penggunaannya." *Malim: Jurnal Pengajian Umum Asia Tenggara (Sea Journal of General Studies)* 21, no. 1 (2020): 160–74. https://doi.org/10.17576/malim-2020-2101-13.

Hendriyani. "Editorial Note: Uji Validitas Dengan Korelasi Item-Total?" *Jurnal Manajemen Strategi Dan Aplikasi Bisnis* 4, no. 1 (2021): 315–20. https://doi.org/https://doi.org/10.36407/jmsab.v4i2.404.

Hilmi, Ahmad Mashadar, Muhammad Syihabul Ihsan Al Haqiqy, and Rahmah Fadhilah Agustina. "Educational Quality Management in Elementary Schools: Measuring the Validity and Reliability of Maharah Qira'ah Test Items." *Journal of Educational Management Research* 4, no. 2 (2025): 931–46. https://doi.org/https://doi.org/10.61987/jemr.v4i2.534.

Karakaya, Sevinc, and Zeliha Alparslan. "Sample Size in Reliability Studies: A Practical Guide Based on Cronbach's Alpha." *Psychiatry and Behavioral Sciences* 12, no. 3 (2022): 150. https://doi.org/10.5455/pbs.20220127074618.

Keith, S. T. "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education." *Research in Science Education* 48 (2018): 1273–96. https://doi.org/10.1007/s11165-016-9602-2.

Kilic, Selim. "Cronbach's Alpha Reliability Coefficient." *Journal of Mood Disorders* 6, no. 1 (2016): 47. https://doi.org/10.5455/jmood.20160307122823.

Lastrijanah, Lastrijanah, Teguh Prasetyo, and Annisa Mawardini. "Pengaruh Media Pembelajaran Geoboard Terhadap Hasil Belajar Siswa." *Didaktika Tauhidi: Jurnal Pendidikan Guru Sekolah Dasar* 4, no. 2 (2017): 87. https://doi.org/10.30997/dt.v4i2.895.

Lazuardi, Fajar, Muhammad Zhafir Al-Hazmi, Ubaid Ridlo, and Raswan. "Pengembangan Instrumen Evaluasi Maharah Qiraah." *As-Sulthan Journal Of Education* 2, no. 4 (2025): 435–424.

Nazilah, Ulfatun, and Rusdiana Navlia. "Evaluasi Program Pendidikan Dalam Meningkatkan Kualitas Pembelajaran." *JIMAD Jurnal Ilmiah Mutiara Pendidikan* 4, no. 1 (2026): 1–14. https://doi.org/10.61404/jimad.v4i1.450.

Rejeki, Sri, Angela Bayu Pertama Sari, Dwi Iswahyuni, Devita Widyaningtyas Yogyanti, Sutanto Sutanto, and Helta Anggia. "Discrimination Index, Difficulty Index, and Distractor Efficiency in MCQs English for Academic Purposes Midterm Test." *Journal of English Language and Pedagogy* 6, no. 1 (2023): 1–11. https://doi.org/https://doi.org/10.36597/jelp.v6i1.14738.

Rezigalla, Assad Ali, Ali Mohammed Elhassan Seid Ahmed Eleragi, Amar Babikir Elhussein, Jaber Alfaifi, Mushabab A. ALGhamdi, Ahmed Y. Al Ameer, Amar Ibrahim Omer Yahia, Osama A. Mohammed, and Masoud Ishag Elkhalifa Adam. "Item Analysis: The Impact of Distractor Efficiency on the Difficulty Index and Discrimination Power of Multiple-Choice Items." *BMC Medical Education* 24, no. 1 (2024): 2–7. https://doi.org/10.1186/s12909-024-05433-y.

Subhaktiyasa Putu Gede. "Evaluasi Validitas Dan Reliabilitas Instrumen Penelitian Kuantitatif: Sebuah Studi Pustaka." *Journal of Education Research* 5 (2024): 5599–5609.

Suharsimi Arikunto. *Dasar-Dasar Evaluasi Pendidikan*. Edited by Restu Damayanti. 3rd ed. Jakarta: PT Bumi Aksara, 2018.

Sumaryanta. *Teori Tes Klasik Dan Teori Respon Butir: Konsep Dan Contoh Penerapannya*. *Cetakan Pertama*. 1st ed. Vol. 15. Cirebon: CV. Confident, 2021.

Suseno, Endro, and Purwo Susongko. *Mengukur Validitas Tes*. Edited by Endro Suseno. 1st ed. Jawa Timur: Pameral Edukreatif, 2021.

Waruwu, Marinu, Siti Natijatul Pu`at, Patrisia Rahayu Utami, Elli Yanti, and Marwah Rusydiana. "Metode

Penelitian Kuantitatif: Konsep, Jenis, Tahapan Dan Kelebihan." *Jurnal Ilmiah Profesi Pendidikan* 10, no. 1 (2025): 917–32. https://doi.org/10.29303/jipp.v10i1.3057.

Yustiandi, and Duden Saepuzaman. "Teori Analisis Butir Soal, Nilai Daya Pembeda Yang Negatif Menunjukkan Bahwa Butir Soal Tidak Mampu Membedakan Peserta Didik Berkemampuan Tinggi Dan Rendah, Bahkan Berpotensi Berlawanan Dengan Tujuan Pengukuran." *Basicedu* 8, no. 6 (2024): 4700–4706. https://doi.org/https://doi.org/10.31004/basicedu.v8i6.9031.

Zaenal, Arifin. "Kriteria Instrumen Dalam Suatu Penelitian." *Jurnal THEOREMS (The Original Research of Mathematics)* 2, no. 1 (2017): 28–36.

Zhang, Sijun, and Kimberly Colvin. "Comparison of Different Reliability Estimation Methods for Single-Item Assessment: A Simulation Study." *Frontiers in Psychology* 15, no. 1 (2024): 148. https://doi.org/10.3389/fpsyg.2024.1482016.