PENGEMBANGAN INSTRUMEN *QIRĀ'AH* BAHASA ARAB YANG VALID DAN RELIABEL

N.F. Hamidah[1], Syahrul[2], Patahuddin[3]

[123]Program Studi Penelitian dan Evaluasi Pendidikan
Universitas Negeri Makassar
Makassar, Indonesia

e-mail: 240012301003@student.unm.ac.id[1], syahrul@unm.ac.id[2], patahuddin1964@gmail.com[3]

Abstrak

Penelitian ini bertujuan mengembangkan dan mengevaluasi kualitas instrumen tes sumatif pilihan ganda untuk mengukur keterampilan membaca (*qirā'ah*) bahasa Arab siswa kelas VIII madrasah tsanawiyah. Penelitian menggunakan pendekatan penelitian dan pengembangan dengan model Borg dan Gall yang disederhanakan menjadi tujuh tahap, yaitu studi pendahuluan, perencanaan, pengembangan produk awal, validasi ahli, uji coba produk, revisi, dan penyempurnaan produk akhir. Sebanyak 100 butir soal dikembangkan dalam dua paket, masing-masing terdiri atas 50 butir. Untuk uji empiris, 20 butir dipilih secara purposif dari Paket I dengan mempertimbangkan keterwakilan indikator dan cakupan materi berdasarkan kisi-kisi, serta durasi pelaksanaan tes. Uji coba melibatkan 162 siswa kelas VIII di MTsN 1 Makassar. Data dianalisis menggunakan indeks Aiken untuk validitas isi, analisis faktor eksploratori (*exploratory factor analysis*/EFA), koefisien alpha Cronbach untuk konsistensi internal, serta analisis tingkat kesukaran dan korelasi item-rest. Nilai Aiken's V yang dilaporkan berkisar antara 0,975 dan 1,00. Analisis EFA menghasilkan nilai Kaiser–Meyer–Olkin (KMO) sebesar 0,714, sedangkan solusi satu faktor yang dipaksakan menjelaskan 17,4% varians. Hasil tersebut belum cukup untuk menyimpulkan bahwa instrumen memiliki struktur unidimensional. Koefisien alpha Cronbach sebesar 0,724 pada 20 butir dan meningkat menjadi 0,764 setelah butir V20 dihapus. Analisis terhadap 19 butir menunjukkan bahwa 18 butir berkategori mudah dan satu butir berkategori sedang. Beberapa butir juga menunjukkan korelasi item-rest yang rendah. Temuan ini memberikan bukti awal mengenai validitas isi dan konsistensi internal instrumen, tetapi revisi butir serta pengujian struktur konstruk lebih lanjut masih diperlukan.

Kata kunci: Bahasa Arab; Instrumen Tes Sumatif; Reliabilitas; Validitas

Abstract

This study aimed to develop and evaluate a multiple-choice summative test instrument for assessing Arabic reading skills (qirā'ah) among Grade VIII students at an Islamic junior high school. The study employed a research and development approach based on a simplified seven-stage Borg and Gall model: preliminary study, planning, initial product development, expert validation, product try-out, revision, and final product refinement. A total of 100 items were developed in two test packages, each consisting of 50 items. For empirical testing, 20 items were purposively selected from Package I based on indicator representation and content coverage in the test blueprint, as well as test administration time considerations. The try-out involved 162 Grade VIII students at MTsN 1 Makassar. Data were analyzed using Aiken's V for content validity, exploratory factor analysis (EFA), Cronbach's alpha for internal consistency, and item difficulty and item-rest correlation analyses. The reported Aiken's V values ranged from 0.975 to 1.00. The EFA yielded a Kaiser–Meyer–Olkin (KMO) value of 0.714, while the forced single-factor solution explained 17.4% of the variance. These findings were insufficient to establish the instrument's unidimensionality. Cronbach's alpha was 0.724 for the 20-item set and increased to 0.764 after item V20 was removed. Analysis of the remaining 19 items classified 18 as easy and one as moderate in difficulty. Several items also exhibited low item-rest correlations. These findings provide preliminary evidence of content validity and internal consistency; however, further item revision and construct validation are needed.

Keywords: *Arabic Language; Summative Test Instrument; Reliability; Validity*

PENDAHULUAN

Bahasa Arab merupakan salah satu mata pelajaran inti di Madrasah Tsanawiyah yang berperan penting dalam memahami sumber ajaran Islam dan literatur keislaman. Penguasaan Bahasa Arab tidak hanya berkaitan dengan kemampuan berkomunikasi, tetapi juga mendukung kemampuan peserta didik dalam memahami informasi dan teks berbahasa Arab. Oleh karena itu, pembelajaran Bahasa Arab perlu didukung oleh sistem evaluasi yang mampu mengukur capaian belajar secara tepat. Evaluasi pembelajaran merupakan proses sistematis untuk menentukan tingkat pencapaian tujuan pembelajaran dan menjadi dasar dalam pengambilan keputusan pendidikan (Arifin, 2012).

Dalam pembelajaran Bahasa Arab, evaluasi mencakup berbagai keterampilan berbahasa, yaitu menyimak, berbicara, membaca, dan menulis. Salah satu keterampilan yang penting untuk diukur adalah membaca (*qirā'ah*), karena keterampilan ini berkaitan dengan kemampuan peserta didik memahami informasi, struktur bahasa, dan makna teks secara menyeluruh (Hartati & Nurhadi, 2021). Kemampuan membaca Bahasa Arab juga melibatkan pemahaman kosakata, struktur kebahasaan, serta kemampuan menginterpretasikan isi teks secara tepat (Ramadhani & Zukhaira, 2019).

Kualitas instrumen evaluasi menentukan ketepatan informasi yang diperoleh dari hasil pengukuran. Instrumen yang baik memerlukan bukti validitas dan reliabilitas agar interpretasi hasil pengukurannya dapat dipertanggungjawabkan (Aiken, 1985). Selain itu, instrumen perlu memiliki karakteristik yang mendukung pengukuran secara objektif dan mampu membedakan kemampuan peserta didik (Laili, 2020). Validitas berkaitan dengan sejauh mana bukti yang tersedia mendukung interpretasi hasil pengukuran sesuai tujuan penggunaannya, sedangkan reliabilitas berkaitan dengan konsistensi hasil pengukuran (Polit & Beck, 2006). Oleh karena itu, pengembangan instrumen pendidikan perlu disertai pengujian empiris untuk memperoleh bukti mengenai kualitas instrumen tersebut (Kamila & Widodo, 2025).

Pengujian kualitas instrumen dapat mencakup validitas isi, analisis struktur konstruk, konsistensi internal, serta karakteristik butir, seperti tingkat kesukaran dan daya pembeda (Dianova & Anwar, 2024). Salah satu metode yang dapat digunakan untuk mengkaji struktur konstruk adalah analisis faktor eksploratori (*exploratory factor analysis* atau EFA), yaitu pendekatan untuk mengeksplorasi pola hubungan antarbutir dalam instrumen pengukuran (Hair et al., 2010). Penerapan EFA dalam pengembangan instrumen pendidikan dapat memberikan informasi mengenai struktur faktor yang didukung oleh data empiris (Hidayah, 2025). Konsistensi internal instrumen juga perlu diperhatikan agar instrumen mampu menghasilkan pengukuran yang stabil dan dapat dipercaya (Rohmah et al., 2023).

Tes sumatif digunakan untuk menilai capaian akhir pembelajaran sehingga instrumen yang digunakan harus mampu mengukur kompetensi peserta didik secara menyeluruh. Dalam konteks keterampilan membaca Bahasa Arab, instrumen tes sumatif dirancang untuk mengukur pemahaman kosakata, struktur kalimat, dan makna teks secara terpadu (Nurdianto, 2020). Namun demikian, pengembangan instrumen tes Bahasa Arab yang memenuhi aspek validitas, reliabilitas, dan kualitas butir soal masih menjadi tantangan dalam praktik pembelajaran.

Beberapa penelitian sebelumnya menunjukkan bahwa pengembangan evaluasi pembelajaran Bahasa Arab terus mengalami perkembangan. Fadhilah mengembangkan penilaian pengetahuan Bahasa Arab berbasis tes objektif untuk meningkatkan kualitas pengukuran kemampuan peserta didik (Fadhilah, 2021). Penelitian Laane menunjukkan bahwa pengembangan tes keterampilan *qirā'ah* dapat mendukung kualitas evaluasi pembelajaran Bahasa Arab secara lebih sistematis (Laane et al., 2024). Penelitian Rahmayanti juga menunjukkan pentingnya pengembangan evaluasi keterampilan berbahasa Arab yang sesuai dengan kebutuhan pembelajaran saat ini (Rahmayanti et al., 2024). Selain itu, penelitian Fitra et al. menunjukkan pentingnya pengujian kualitas instrumen melalui analisis validitas dan karakteristik butir agar instrumen benar-benar layak digunakan dalam proses evaluasi pembelajaran (Fitra et al., 2024).

Meskipun penelitian terkait pengembangan evaluasi Bahasa Arab telah banyak dilakukan, penelitian yang secara khusus mengembangkan instrumen tes sumatif keterampilan membaca Bahasa Arab dengan mengintegrasikan validitas isi, validitas konstruk, reliabilitas, serta analisis karakteristik butir pada jenjang Madrasah Tsanawiyah masih relatif terbatas. Oleh karena itu, penelitian ini bertujuan mengembangkan instrumen tes sumatif keterampilan membaca Bahasa Arab serta memperoleh bukti awal mengenai validitas isi, struktur faktor, konsistensi internal, dan karakteristik butir pada siswa kelas VIII di MTsN 1 Makassar.

METODE

Penelitian ini menggunakan pendekatan penelitian dan pengembangan (*Research and Development* atau R&D) dengan mengadaptasi model pengembangan Borg dan Gall (1983) menjadi tujuh tahap, yaitu studi pendahuluan, perencanaan, pengembangan produk awal, validasi ahli, uji coba produk, revisi produk, dan penyempurnaan produk akhir. Model tersebut disederhanakan dengan memusatkan proses pengembangan pada tahapan yang relevan dengan tujuan penelitian, yaitu menghasilkan instrumen tes sumatif pilihan ganda untuk mengukur keterampilan membaca (*qirā'ah*) Bahasa Arab. Prosedur penelitian mencakup kajian kebutuhan dan teori, penyusunan instrumen, validasi ahli, pengujian empiris, serta peninjauan kualitas butir berdasarkan hasil analisis (Borg & Gall, 1983).

Pada tahap studi pendahuluan, dilakukan kajian literatur, observasi awal, dan wawancara dengan guru Bahasa Arab kelas VIII di MTsN 1 Makassar. Hasil studi pendahuluan menunjukkan bahwa evaluasi yang digunakan masih cenderung berfokus pada hafalan kosakata dan penerjemahan literal, sehingga belum sepenuhnya mencakup pemahaman struktur kalimat dan makna teks secara menyeluruh. Temuan tersebut menjadi dasar penetapan tiga aspek keterampilan membaca sebagai acuan pengembangan instrumen, yaitu pemahaman kosakata (mufradāt), pemahaman struktur kalimat (tarkīb), dan pemahaman makna teks. Pada tahap perencanaan, aspek tersebut dijabarkan ke dalam kisi-kisi yang memuat kompetensi atau tujuan pembelajaran, materi, indikator soal, tingkat kognitif, dan nomor butir. Kisi-kisi mencakup enam tema bacaan, yaitu At-Ta'āruf (perkenalan), Al-Marāfiq al-Madrasiyyah (fasilitas sekolah), Al-'Adawāt al-Madrasiyyah (peralatan sekolah), An-Naẓāfah (kebersihan), Al-Usrah (keluarga), dan Al-Waqtu (waktu). Pemetaan 20 butir yang dipilih untuk uji coba empiris berdasarkan tema, indikator, dan tingkat kognitif disajikan pada Tabel 1 di bagian akhir manuskrip.

Tahap pengembangan produk awal menghasilkan 100 butir soal pilihan ganda berbasis teks yang dibagi menjadi dua paket yang dirancang paralel, masing-masing terdiri atas 50 butir. Setiap butir terdiri atas satu stem dan empat alternatif jawaban, yaitu satu jawaban benar dan tiga pengecoh. Penyusunan butir memperhatikan kesesuaian materi dengan indikator keterampilan membaca, kejelasan stem, konstruksi alternatif jawaban, dan ketepatan penggunaan bahasa Arab sesuai dengan karakteristik siswa kelas VIII. Kedua paket dirancang dengan cakupan materi, indikator, dan tingkat kognitif yang sepadan. Namun, karena uji coba empiris hanya dilakukan terhadap butir dari Paket I, kesetaraan psikometrik kedua paket belum dapat dikonfirmasi melalui penelitian ini.

Validasi ahli dilakukan oleh empat validator, yaitu dua ahli Bahasa Arab, satu ahli pengukuran dan penilaian pendidikan, serta satu ahli linguistik. Validator menilai kesesuaian materi dengan indikator keterampilan membaca, konstruksi soal, kejelasan alternatif jawaban, dan ketepatan penggunaan bahasa. Penilaian dilakukan menggunakan lembar validasi dengan skala numerik 1–5 serta ruang komentar untuk masukan kualitatif. Validitas isi dianalisis menggunakan indeks Aiken's V untuk mengetahui tingkat kesesuaian butir berdasarkan penilaian ahli (Aiken, 1985). Dalam penelitian ini, butir dinyatakan memenuhi kriteria kelayakan isi apabila memperoleh nilai Aiken's V ≥ 0,80, sesuai dengan kriteria yang digunakan dalam penelitian.

Dari 100 butir yang dikembangkan, peneliti memilih 20 butir dari Paket I untuk uji coba empiris secara purposif. Pemilihan mempertimbangkan keterwakilan tema dan indikator berdasarkan kisi-kisi serta cakupan materi yang diujikan. Jumlah butir yang dipilih juga

mempertimbangkan masukan guru Bahasa Arab yang menilai bahwa pelaksanaan uji coba dengan 50 butir sekaligus terlalu panjang. Dua puluh butir yang dipilih adalah nomor 1, 5, 7, 9, 12, 15, 17, 20, 24, 25, 29, 31, 33, 36, 39, 40, 41, 44, 47, dan 50 pada Paket I. Butir-butir tersebut mencakup keenam tema bacaan dalam kisi-kisi. Dengan demikian, hasil pengujian empiris secara langsung menggambarkan karakteristik 20 butir yang diujicobakan dan tidak dapat langsung digeneralisasikan kepada seluruh 100 butir dalam kedua paket.

Uji coba dilakukan terhadap 162 siswa kelas VIII di MTsN 1 Makassar. Pemilihan subjek dilakukan secara purposif berdasarkan kesesuaian dengan konteks pembelajaran yang menjadi sasaran pengembangan instrumen. Data penelitian dikumpulkan melalui lembar validasi ahli dan respons siswa terhadap 20 butir soal. Respons siswa dikodekan secara dikotomis, yaitu 1 untuk jawaban benar dan 0 untuk jawaban salah. Data penilaian ahli digunakan untuk menganalisis validitas isi, sedangkan respons siswa digunakan untuk mengkaji struktur faktor, konsistensi internal, dan karakteristik butir.

Pengolahan data statistik dilakukan menggunakan perangkat lunak JASP (Goss-Sampson, 2019). Validitas konstruk dianalisis menggunakan analisis faktor eksploratori (*exploratory factor analysis* atau EFA) untuk mengeksplorasi struktur hubungan antarbutir dalam instrumen (Hair et al., 2010). Penggunaan analisis faktor eksploratori dalam pengembangan instrumen pendidikan juga telah digunakan untuk memperoleh bukti mengenai struktur konstruk (Hidayah, 2025). Berdasarkan prosedur analisis yang dilaporkan dalam penelitian ini, EFA menggunakan metode *Principal Axis Factoring* (PAF) dengan matriks korelasi biasa dan rotasi Oblimin. Kelayakan data diperiksa melalui Kaiser–Meyer–Olkin (KMO) dan Bartlett's Test of Sphericity. Interpretasi hasil mempertimbangkan nilai *factor loading*, jumlah faktor yang didukung oleh data, dan proporsi varians yang dijelaskan. Nilai *factor loading* sebesar 0,30 digunakan sebagai acuan awal untuk meninjau kontribusi butir, bukan sebagai satu-satunya dasar untuk menetapkan validitas konstruk. Karena data respons bersifat dikotomis dan analisis awal menggunakan matriks korelasi biasa, hasil EFA perlu ditafsirkan secara hati-hati. Reliabilitas instrumen dihitung menggunakan koefisien Cronbach's Alpha untuk mengkaji konsistensi internal antarbutir (Rohmah et al., 2023).

Analisis awal dilakukan terhadap 20 butir hasil uji coba. Selanjutnya, analisis reliabilitas dihitung kembali setelah V20 dikeluarkan karena butir tersebut menunjukkan korelasi item-rest negatif. Penghitungan ulang dilakukan terhadap 19 butir untuk mengevaluasi perubahan konsistensi internal setelah penghapusan V20. Hasil analisis untuk 20 dan 19 butir dilaporkan secara terpisah agar jumlah butir yang digunakan pada setiap analisis dapat dibedakan dengan jelas. Karakteristik butir dianalisis menggunakan pendekatan teori tes klasik (*Classical Test Theory* atau CTT), meliputi tingkat kesukaran dan daya pembeda (Baker, 2001).

Tingkat kesukaran dihitung berdasarkan proporsi siswa yang menjawab benar pada setiap butir, sedangkan daya pembeda ditinjau melalui korelasi item-rest, yaitu hubungan antara skor suatu butir dan skor total tanpa memasukkan butir tersebut. Hasil analisis digunakan untuk mengidentifikasi butir yang perlu dipertahankan, ditinjau kembali, atau dikeluarkan dari analisis lanjutan. Teori respons butir merupakan pendekatan pengukuran lain yang berbeda dari CTT, tetapi tidak digunakan dalam analisis empiris penelitian ini. Efektivitas pengecoh tidak dianalisis berdasarkan frekuensi pilihan setiap opsi karena data yang digunakan dalam analisis berupa skor benar-salah, bukan data pilihan alternatif jawaban siswa (DeMars, 2018).

HASIL DAN PEMBAHASAN

Instrumen yang dikembangkan merupakan tes sumatif pilihan ganda untuk mengukur keterampilan membaca Bahasa Arab siswa kelas VIII. Instrumen awal terdiri atas 100 butir yang dibagi menjadi dua paket yang dirancang paralel, masing-masing sebanyak 50 butir. Uji coba empiris dilakukan terhadap 20 butir yang dipilih dari Paket I dan diberikan kepada 162 siswa kelas VIII di MTsN 1 Makassar. Pemilihan butir dilakukan secara purposif dengan

mempertimbangkan keterwakilan tema dan indikator pada kisi-kisi serta masukan guru mengenai jumlah soal yang dapat dilaksanakan dalam satu kali uji coba. Pemetaan 20 butir yang diujicobakan disajikan pada Tabel 1.

Tabel 1. Kisi-Kisi 20 Butir Instrumen Qirā'ah yang Diujicobakan

Kode butir	Nomor butir awal Paket I	Tema dan kompetensi/tujuan pembelajaran	Level kognitif	Indikator soal
V1	1	Tema 1: At-Ta'āruf — memahami dan menafsirkan teks tentang perkenalan diri dan kegiatan di sekolah	C1	Menentukan nama tokoh dalam teks
V2	5	Tema 1: At-Ta'āruf — memahami dan menafsirkan teks tentang perkenalan diri dan kegiatan di sekolah	C3	Menentukan kegiatan yang dilakukan di kelas
V3	7	Tema 1: At-Ta'āruf — memahami dan menafsirkan teks tentang perkenalan diri dan kegiatan di sekolah	C5	Menilai sikap belajar dari isi teks
V4	9	Tema 2: Al-Marāfiq al-Madrasiyyah — memahami teks deskriptif tentang fasilitas dan lingkungan sekolah	C1	Menentukan nama fasilitas dalam teks
V5	12	Tema 2: Al-Marāfiq al-Madrasiyyah — memahami teks deskriptif tentang fasilitas dan lingkungan sekolah	C2	Menentukan kegiatan setelah belajar
V6	15	Tema 2: Al-Marāfiq al-Madrasiyyah — memahami teks deskriptif tentang fasilitas dan lingkungan sekolah	C5	Mengevaluasi pesan moral tentang kebersihan
V7	17	Tema 3: Al-'Adawāt al-Madrasiyyah — memahami teks tentang peralatan sekolah dan pemeliharaannya	C1	Menentukan isi tas siswa
V8	20	Tema 3: Al-'Adawāt al-Madrasiyyah — memahami teks tentang peralatan sekolah dan pemeliharaannya	C2	Memahami alasan menjaga perlengkapan
V9	24	Tema 3: Al-'Adawāt al-Madrasiyyah — memahami teks tentang peralatan sekolah dan pemeliharaannya	C5	Mengevaluasi akibat jika tidak merawat alat
V10	25	Tema 4: An-Naẓāfah — memahami dan menilai teks tentang kebersihan diri, kelas, dan lingkungan sekolah	C1	Menafsirkan arti "kebersihan sebagian dari iman"

Kode butir	Nomor butir awal Paket I	Tema dan kompetensi/tujuan pembelajaran	Level kognitif	Indikator soal
V11	29	Tema 4: An-Naẓāfah — memahami dan menilai teks tentang kebersihan diri, kelas, dan lingkungan sekolah	C3	Menerapkan pemahaman kebersihan
V12	31	Tema 4: An-Naẓāfah — memahami dan menilai teks tentang kebersihan diri, kelas, dan lingkungan sekolah	C5	Mengevaluasi nilai moral dalam teks
V13	33	Tema 5: Al-Usrah — memahami teks tentang anggota keluarga dan peran mereka	C1	Menentukan jumlah anggota keluarga
V14	36	Tema 5: Al-Usrah — memahami teks tentang anggota keluarga dan peran mereka	C2	Menentukan waktu berkumpul
V15	39	Tema 5: Al-Usrah — memahami teks tentang anggota keluarga dan peran mereka	C5	Mengevaluasi pentingnya menghormati keluarga
V16	40	Tema 5: Al-Usrah — memahami teks tentang anggota keluarga dan peran mereka	C5	Mengevaluasi alasan keluarga mengajarkan tanggung jawab
V17	41	Tema 6: Al-Waqtu — memahami urutan waktu, kegiatan, dan manajemen waktu sehari-hari	C1	Menentukan waktu bangun
V18	44	Tema 6: Al-Waqtu — memahami urutan waktu, kegiatan, dan manajemen waktu sehari-hari	C2	Menentukan waktu pulang sekolah
V19	47	Tema 6: Al-Waqtu — memahami urutan waktu, kegiatan, dan manajemen waktu sehari-hari	C5	Mengevaluasi kebiasaan mengatur waktu
V20	50	Tema 6: Al-Waqtu — memahami urutan waktu, kegiatan, dan manajemen waktu sehari-hari	C5	Mengevaluasi akibat tidak mengatur waktu

Validitas isi diperoleh melalui penilaian empat validator, yang terdiri atas dua ahli Bahasa Arab, satu ahli pengukuran dan penilaian pendidikan, serta satu ahli linguistik. Penilaian mencakup kesesuaian materi, konstruksi butir, dan penggunaan bahasa. Berdasarkan perhitungan Aiken's V terhadap 100 butir yang dikembangkan, nilai indeks berada pada rentang 0.975–1.000. Untuk 20 butir yang diujicobakan secara empiris, nilai Aiken's V berada pada rentang 0.981–0.995. Rincian nilai Aiken's V untuk butir yang diuji disajikan pada Tabel 2.

Tabel 2. Nilai Aiken's V pada 20 Butir yang Diujicobakan

Kode butir	Nomor butir awal Paket I	Aiken's V
V1	1	0.9828125
V2	5	0.9890625
V3	7	0.9812500
V4	9	0.9937500
V5	12	0.9906250
V6	15	0.9906250
V7	17	0.9875000
V8	20	0.9953125
V9	24	0.9921875
V10	25	0.9921875
V11	29	0.9906250
V12	31	0.9937500
V13	33	0.9953125
V14	36	0.9859375
V15	39	0.9953125
V16	40	0.9953125
V17	41	0.9953125
V18	44	0.9937500
V19	47	0.9937500
V20	50	0.9953125

Rentang nilai tersebut menunjukkan bahwa penilaian para ahli terhadap kesesuaian butir dengan kriteria validasi cenderung tinggi. Meskipun demikian, bukti validitas isi perlu dibedakan dari bukti psikometrik yang diperoleh melalui respons siswa. Nilai Aiken's V yang tinggi tidak dengan sendirinya membuktikan bahwa setiap butir memiliki struktur faktor, daya pembeda, atau tingkat kesukaran yang memadai. Oleh sebab itu, hasil validitas isi ditafsirkan bersama-sama dengan hasil analisis empiris pada tahap berikutnya.

Validitas konstruk dikaji menggunakan analisis faktor eksploratori (*Exploratory factor analysis* atau EFA) terhadap respons 162 siswa pada 20 butir soal. Hasil pemeriksaan kelayakan data menunjukkan nilai KMO keseluruhan sebesar 0.714 dan hasil *Bartlett's Test of Sphericity* sebesar $\chi^2 = 806.370$, $df = 190$, $p < 0.001$. Temuan tersebut menunjukkan bahwa pola korelasi antarbutir secara keseluruhan memungkinkan dilakukannya eksplorasi faktor. Nilai ini merupakan informasi mengenai kelayakan matriks untuk analisis faktor, bukan bukti yang cukup untuk menyatakan bahwa instrumen telah memiliki validitas konstruk yang kuat. Hasil analisis faktor eksploratori disajikan pada Tabel 3.

Tabel 3. Output KMO dan Bartlett's Test

KMO / Overall MSA	0.714
Bartlett's test / χ^2	806.370
Bartlett's test / df	190.000
Bartlett's test / p	< 0.001

Analisis awal dalam penelitian ini menggunakan metode *Principal Axis Factoring* (PAF) dengan matriks korelasi biasa dan spesifikasi satu faktor. Berdasarkan output yang tersedia, 13 butir menampilkan *factor loading* sebesar 0.313–0.780. Nilai tertinggi ditemukan pada V17 (0.780), diikuti V15 (0.690), V7 (0.581), V16 (0.579), dan V18 (0.579). Sementara itu, nilai *factor loading* untuk V1, V3, V4, V5, V6, V10, dan V20 tidak ditampilkan dalam tabel output yang tersedia. Ketujuh butir tersebut juga memiliki nilai *uniqueness* yang tinggi, yaitu 0.915–0.996. Dengan demikian, kontribusi butir-butir tersebut terhadap faktor

yang ditetapkan perlu ditafsirkan secara hati-hati. Hasil *factor loading* dan *uniqueness* disajikan pada Tabel 4.

Tabel 4. Hasil Factor Loading dan Uniqueness EFA

Kode butir	Nomor butir awal Paket I	Factor 1	Uniqueness
V1	1		0.934
V2	5	0.313	0.902
V3	7		0.986
V4	9		0.915
V5	12		0.966
V6	15		0.963
V7	17	0.581	0.662
V8	20	0.325	0.895
V9	24	0.358	0.872
V10	25		0.978
V11	29	0.415	0.828
V12	31	0.446	0.801
V13	33	0.415	0.828
V14	36	0.349	0.878
V15	39	0.690	0.524
V16	40	0.579	0.665
V17	41	0.780	0.392
V18	44	0.579	0.664
V19	47	0.363	0.868
V20	50		0.996

Faktor pertama memiliki *eigenvalue* sebesar 3.484 dan menjelaskan 17.4% varians. Walaupun nilai *eigenvalue* melebihi satu, proporsi varians yang dijelaskan relatif terbatas. Selain itu, hasil *parallel analysis* pada analisis awal mengindikasikan kemungkinan terbentuknya lebih dari satu faktor. Karena solusi satu faktor telah ditetapkan dalam analisis, hasil tersebut belum cukup untuk memastikan bahwa struktur instrumen benar-benar unidimensional. Temuan EFA lebih tepat dipandang sebagai bukti eksploratif awal mengenai struktur faktor, dengan keterbatasan pada kontribusi sejumlah butir dan proporsi varians yang dijelaskan. Ringkasan karakteristik faktor disajikan pada Tabel 5.

Tabel 5. Eigenvalue dan Proporsi Varians Faktor

Factor	Eigenvalue	Proportion var.	Cumulative
Factor 1	3.484	0.174	0.174

Interpretasi ini juga perlu mempertimbangkan bahwa respons butir berbentuk dikotomis, sedangkan analisis awal menggunakan matriks korelasi biasa. Dengan demikian, hasil EFA yang tersedia belum memberikan gambaran dari analisis berbasis korelasi tetrakorik. Temuan struktur faktor dalam penelitian ini tidak digunakan sebagai satu-satunya dasar untuk menyatakan seluruh butir valid secara konstruk.

Reliabilitas dianalisis menggunakan koefisien Cronbach's Alpha. Pada analisis awal terhadap 20 butir, diperoleh nilai alpha sebesar 0.724 dengan interval kepercayaan 95% sebesar 0.614–0.835. Pemeriksaan korelasi *item-rest* menunjukkan bahwa V20 memiliki nilai negatif, yaitu −0.122. Nilai tersebut menunjukkan bahwa respons terhadap V20 bergerak berlawanan dengan pola skor pada butir lainnya dalam perhitungan tersebut.

Setelah V20 dikeluarkan dari penghitungan ulang, nilai Cronbach's Alpha untuk 19 butir meningkat menjadi 0.764 dengan interval kepercayaan 95% sebesar 0.670–0.859. Peningkatan ini menunjukkan perubahan konsistensi internal setelah V20 dikeluarkan dari analisis reliabilitas. Namun, kenaikan alpha tidak dengan sendirinya membuktikan bahwa

seluruh 19 butir yang tersisa telah memenuhi semua kriteria validitas konstruk. Hasil ini perlu dipertimbangkan bersama dengan temuan EFA dan analisis karakteristik setiap butir. Ringkasan reliabilitas untuk 20 dan 19 butir disajikan pada Tabel 6.

Tabel 6. Ringkasan Reliabilitas dan Evaluasi Butir V20

Analisis	Jumlah butir	Estimate	Lower 95% CI	Upper 95% CI
Cronbach's alpha sebelum pengeluaran butir	20	0.724	0.614	0.835
Cronbach's alpha setelah V20 dikeluarkan	19	0.764	0.670	0.859
Item-rest correlation V20 pada analisis 20 butir	20	-0.122	-0.271	0.033

Analisis karakteristik butir setelah V20 dikeluarkan dilakukan terhadap 19 butir. Tingkat kesukaran dihitung berdasarkan proporsi jawaban benar (p). Hasil analisis menunjukkan bahwa 18 dari 19 butir berada pada kategori mudah, sedangkan satu butir, yaitu V5, berada pada kategori sedang dengan nilai p sebesar 0.691. Tidak terdapat butir yang berada pada kategori sukar. Enam butir memiliki proporsi jawaban benar di atas 0.90, yaitu V1 (0.963), V4 (0.975), V6 (0.920), V7 (0.932), V8 (0.951), dan V19 (0.938). Hasil tersebut menunjukkan bahwa sebaran tingkat kesukaran belum seimbang dan sebagian besar butir relatif mudah bagi responden uji coba.

Butir yang sangat mudah tidak otomatis tidak layak, karena interpretasinya juga bergantung pada tujuan tes. Namun, proporsi jawaban benar yang tinggi mengurangi kemampuan suatu butir untuk membedakan peserta dengan tingkat kemampuan yang berbeda apabila hampir semua siswa menjawab benar. Oleh sebab itu, butir dengan p di atas 0.90 perlu ditinjau kembali dari sisi tuntutan pemahaman bacaan, ketepatan kunci jawaban, dan kesesuaian alternatif jawaban sebelum digunakan kembali. Data yang tersedia menunjukkan perlunya peninjauan tersebut, bukan bahwa revisi butir telah dilakukan setelah uji coba.

Daya pembeda ditinjau melalui korelasi *item-rest*. Beberapa butir memerlukan perhatian khusus. V3 memiliki korelasi 0.138, V6 sebesar 0.185, dan V10 sebesar 0.165; ketiganya berada di bawah 0.20. V5 memiliki nilai 0.203 dan V19 sebesar 0.230, sehingga meskipun korelasinya positif, kontribusinya relatif terbatas dibandingkan butir dengan korelasi lebih tinggi. Temuan ini mendukung perlunya peninjauan isi dan kesesuaian indikator pada butir-butir tersebut. Namun, keputusan mempertahankan, merevisi, atau menghapus butir tidak sebaiknya didasarkan pada satu indeks saja, melainkan mempertimbangkan kesesuaian isi, hasil validasi ahli, pola *factor loading*, tingkat kesukaran, dan tujuan penggunaan tes.

Ringkasan tingkat kesukaran dan korelasi *item-rest* untuk 19 butir disajikan pada Tabel 7. Pelaporan kedua indeks secara berdampingan membantu menunjukkan bahwa butir yang mudah belum tentu memiliki daya pembeda yang baik, sehingga interpretasi kualitas butir perlu dilakukan secara menyeluruh.

Tabel 7. Tingkat Kesukaran dan Item-Rest Correlation pada 19 Butir

Kode butir	Nomor butir awal Paket I	Proporsi benar (p)	Kategori kesukaran	*Item-rest correlation*	Lower 95% CI	Upper 95% CI
V1	1	0.963	Mudah	0.265	0.116	0.403
V2	5	0.895	Mudah	0.276	0.127	0.413
V3	7	0.870	Mudah	0.138	-0.017	0.286
V4	9	0.975	Mudah	0.292	0.144	0.427
V5	12	0.691	Sedang	0.203	0.051	0.346
V6	15	0.920	Mudah	0.185	0.031	0.329

Kode butir	Nomor butir awal Paket I	Proporsi benar (p)	Kategori kesukaran	*Item-rest correlation*	Lower 95% CI	Upper 95% CI
V7	17	0.932	Mudah	0.462	0.355	0.593
V8	20	0.951	Mudah	0.351	0.208	0.479
V9	24	0.877	Mudah	0.287	0.139	0.422
V10	25	0.747	Mudah	0.165	0.011	0.311
V11	29	0.759	Mudah	0.455	0.323	0.569
V12	31	0.877	Mudah	0.383	0.244	0.508
V13	33	0.852	Mudah	0.383	0.243	0.507
V14	36	0.846	Mudah	0.286	0.138	0.422
V15	39	0.864	Mudah	0.541	0.422	0.642
V16	40	0.796	Mudah	0.485	0.359	0.596
V17	41	0.877	Mudah	0.621	0.519	0.709
V18	44	0.821	Mudah	0.481	0.352	0.591
V19	47	0.938	Mudah	0.230	0.078	0.371

Penelitian ini memiliki beberapa keterbatasan. Pertama, uji empiris dilakukan pada 162 siswa kelas VIII di satu madrasah, yaitu MTsN 1 Makassar, sehingga hasilnya belum dapat digeneralisasikan secara langsung kepada siswa MTs di sekolah atau wilayah lain. Kedua, pengujian empiris hanya mencakup 20 butir dari Paket I. Oleh karena itu, karakteristik psikometrik Paket II dan keseluruhan 100 butir belum dapat dipastikan melalui uji empiris penelitian ini.

Ketiga, hasil EFA perlu ditafsirkan secara terbatas karena sejumlah butir tidak menampilkan *factor loading* dalam output yang tersedia, varians yang dijelaskan oleh faktor pertama sebesar 17.4%, dan *parallel analysis* mengindikasikan kemungkinan lebih dari satu faktor. Analisis awal juga menggunakan matriks korelasi biasa pada data dikotomis. Keempat, respons siswa yang tersedia untuk analisis butir berbentuk skor benar-salah, bukan distribusi pilihan A, B, C, dan D. Oleh karena itu, efektivitas setiap pengecoh tidak dapat ditentukan secara empiris dari data tersebut. Penelitian ini juga tidak menyediakan bukti kuantitatif yang memadai untuk menyimpulkan bahwa instrumen praktis dan efektif dalam pengertian yang lebih luas.

Secara keseluruhan, penelitian ini memberikan bukti validitas isi yang tinggi berdasarkan penilaian ahli dan informasi awal mengenai konsistensi internal serta karakteristik butir. Akan tetapi, bukti struktur faktor dan kualitas butir menunjukkan bahwa sejumlah aspek masih perlu ditinjau. Hasil penelitian karena itu lebih tepat digunakan untuk menjelaskan karakteristik butir yang telah diujicobakan dan mengidentifikasi bagian yang memerlukan perbaikan, bukan untuk menyimpulkan bahwa seluruh instrumen telah terbukti unidimensional, sepenuhnya valid, dan siap digeneralisasikan ke berbagai konteks madrasah.

PENUTUP

Berdasarkan hasil penelitian, instrumen tes sumatif keterampilan membaca Bahasa Arab yang dikembangkan memiliki bukti awal validitas isi dan reliabilitas yang memadai, tetapi masih memerlukan penyempurnaan pada beberapa aspek. Validitas isi berdasarkan penilaian empat validator menunjukkan nilai Aiken's V pada rentang 0,975–1,000 untuk 100 butir yang dikembangkan. Pada pengujian empiris terhadap 20 butir yang dipilih dari Paket I pada 162 siswa kelas VIII di satu MTsN di Kota Makassar, nilai KMO sebesar 0,714 dan Bartlett's test yang signifikan ($p<0,001$) menunjukkan bahwa data memenuhi indikasi awal untuk analisis faktor. Namun, hasil EFA belum memberikan bukti yang cukup kuat untuk menyimpulkan bahwa instrumen bersifat unidimensional, karena faktor pertama menjelaskan 17,4% varians dan sejumlah butir tidak memiliki nilai *factor loading* yang ditampilkan dalam output analisis.

Reliabilitas awal 20 butir menunjukkan nilai Cronbach's Alpha sebesar 0,724. Setelah butir V20 yang memiliki *item-rest correlation* negatif dikeluarkan, nilai alpha meningkat

menjadi 0,764 pada 19 butir. Meskipun demikian, beberapa butir masih perlu ditinjau lebih lanjut karena memiliki *item-rest correlation* rendah. Analisis tingkat kesukaran juga menunjukkan bahwa 18 dari 19 butir berada pada kategori mudah, sedangkan satu butir berada pada kategori sedang. Temuan tersebut menunjukkan bahwa revisi masih diperlukan, khususnya untuk memperbaiki variasi tingkat kesukaran dan mengevaluasi kontribusi butir terhadap skor total. Dengan demikian, instrumen ini memiliki potensi sebagai alat evaluasi sumatif keterampilan membaca Bahasa Arab, tetapi bukti yang tersedia belum cukup untuk menyatakan bahwa seluruh butir telah memenuhi kualitas pengukuran yang diharapkan. Penelitian ini juga belum menyediakan bukti empiris yang memadai untuk menyimpulkan efektivitas dan kepraktisan instrumen, serta belum menguji kesetaraan kedua paket soal secara empiris.

Instrumen yang dikembangkan diharapkan dapat menjadi salah satu alternatif dalam penyusunan evaluasi sumatif keterampilan membaca Bahasa Arab. Namun, pemanfaatannya perlu mempertimbangkan hasil evaluasi butir dan konteks penggunaannya, terutama karena uji empiris dilakukan pada siswa kelas VIII di satu MTsN di Kota Makassar. Hasil penelitian ini belum dapat digeneralisasikan secara luas ke madrasah dengan karakteristik peserta didik yang berbeda.

Penelitian selanjutnya disarankan untuk merevisi dan menguji kembali butir yang memiliki *item-rest correlation* rendah atau tingkat kesukaran yang terlalu tinggi, serta mengevaluasi kembali struktur faktor instrumen. Analisis lanjutan dengan mempertimbangkan karakteristik data dikotomis juga diperlukan untuk memperoleh bukti validitas konstruk yang lebih kuat. Selain itu, uji coba pada madrasah dengan karakteristik yang beragam, pengujian empiris terhadap kedua paket soal, serta pengumpulan data khusus mengenai kepraktisan dan efektivitas diperlukan sebelum instrumen digunakan secara lebih luas.

DAFTAR RUJUKAN

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Arifin, Z. (2012). *Evaluasi pembelajaran*. Remaja Rosdakarya.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC.

Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction* (4th ed.). Longman.

DeMars, C. (2018). *Item response theory*. Oxford University Press.

Dianova, F. R., & Anwar, N. (2024). Analisis butir uji validitas, reliabilitas, tingkat kesukaran, dan daya pembeda soal sumatif bahasa Arab SD Islam. *Jurnal Bahasa Daerah Indonesia, 1*(3), 1–13. https://doi.org/10.47134/jbdi.v1i3.2863

Fadhilah, R. A. (2021). Pengembangan penilaian pengetahuan bahasa Arab berbasis tes objektif. *Uktub: Journal of Arabic Studies, 1*(2), 134–142. https://doi.org/10.32678/uktub.v1i2.5817

Fitra, M. N., Chamidah, D., Basit, S. N. R., & Sardiyanah. (2024). Penyusunan butir soal unsur bahasa Arab (bunyi, kosakata, dan struktur kalimat) dan uji instrumen butir soal di MTs S Aluppang. *Naskhi: Jurnal Kajian Pendidikan dan Bahasa Arab, 6*(2), 23–37. https://doi.org/10.47435/naskhi.v6i2.2964

Goss-Sampson, M. A. (2019). *Statistical analysis in JASP: A guide for students*. JASP. https://doi.org/10.6084/m9.figshare.9980744

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*

(7th ed.). Pearson.

Hartati, M., & Nurhadi. (2021). *Pengembangan instrumen penilaian berbasis literasi membaca*. Direktorat SMA.

Hidayah, N. (2025). The validity of student interest instrument in learning Arabic using *exploratory factor analysis*. *Journal of Practice Learning and Educational Development, 5*(1), 104–113. https://doi.org/10.58737/jpled.v5i1.422

Kamila, A., & Widodo, P. B. (2025). Validitas konstruk dalam psikometri: Peran analisis faktor eksploratori dan konfirmatori. *Jurnal Psikologi Pendidikan dan Pengembangan SDM, 13*(1), 1–7. https://doi.org/10.37721/psi.v13i1.1592

Laane, N. A. H., Hula, I. R. N., & Hairudin, H. (2024). Pengembangan tes keterampilan qira'ah menggunakan aplikasi Candy CBT. *Al-Af'idah: Jurnal Pendidikan dan Pengajaran Bahasa Arab, 8*(2), 429–440. https://doi.org/10.52266/al-afidah.v8i2.3401

Laili, M. (2020). Ketepatan konstruksi butir pilihan ganda bahasa Arab. *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, dan Budaya Arab, 3*(2), 111–124. https://doi.org/10.17509/alsuniyat.v3i2.25272

Nurdianto, T. (2020). *Kompetensi dasar pembelajaran bahasa Arab: Common European framework of reference (CEFR)*. Zahir Publishing.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Rahmayanti, I., Muhamad, S., La Udin, Y., Nashihah, D., & Qomari, N. (2024). Pengembangan formative sebagai media evaluasi maharah al-istima' dan al-qira'ah siswa kelas X MAN Kota Batu Malang. *Mantiqu Tayr: Journal of Arabic Language, 4*(1), 257–278. https://doi.org/10.25217/mantiqutayr.v4i1.4134

Ramadhani, A. I., & Zukhaira. (2019). Pengembangan Arabic Treasure Hunter untuk keterampilan membaca bahasa Arab siswa kelas VIII MTs di Kabupaten Klaten. *LISANUL ARAB: Journal of Arabic Learning and Teaching, 8*(1), 74–83. https://doi.org/10.15294/la.v8i1.32548

Rohmah, F., Rusilowati, A., Supriyadi, S., & Widowati, T. (2023). Validitas dan reliabilitas instrumen penilaian berbasis Android. *Jurnal Penelitian Pendidikan Indonesia, 9*(2), 577–585. https://doi.org/10.29210/020232131