EDUCATIONAL JOURNAL
Homepage https://indojurnal.com/index.php/edu
eISSN 3110-1151 & pISSN 3110-1526
Vol. 2, No. 2, Tahun 2026
doi.org/10.63822/2cnbpb11
Hal. 516-527

Tahapan Pengembangan Instrumen Evaluasi Maharah *Kalam* Berdasarkan Kajian Teoretis dan Hasil Penelitian Terdahulu

Sutan Sulaiman[1], Alif Alvian Mawaddi Hamid[2]
Program Magister Pendidikan Bahasa Arab, UIN Syarif Hidayatullah Jakarta[1,2]

*Email sulaimansutanasution@gmail.com; alvinmawaddi@gmail.com

Diterima: 01-11-2026 | Disetujui: 05-11-2027 | Diterbitkan: 07-11-2026

ABSTRACT

*Evaluating *maharah kalam* (speaking skills) requires instruments capable of measuring speaking proficiency in a valid and reliable manner; however, current practices often rely on assessing comprehension without the use of standardized rubrics. This study aims to outline the stages for developing *maharah kalam* evaluation instruments, drawing upon theoretical frameworks and previous research. Employing a qualitative descriptive approach through a literature review, this study examines theories of language assessment, instrument development models (such as those by Mardapi, the 4D and 3D models, and the ADDIE model), as well as ten studies published between 2021 and 2026. The findings indicate that while existing research typically involves needs analysis, the design of tasks and rubrics, and expert validation, evidence of reliability—particularly inter-rater reliability—is rarely reported. Based on this synthesis, seven development stages are proposed: (1) needs analysis and construct definition; (2) specification and blueprint development; (3) design of tasks and rubrics; (4) expert validation; (5) limited pilot testing; (6) empirical analysis; and (7) revision, final assembly, and reporting. This framework is intended to serve as a guide for educators and researchers in developing systematic and accountable maharah kalam evaluation instruments.*

Keywords: *maharah kalam; evaluation instrument; development stages; assessment rubric; literature review*

ABSTRAK

Evaluasi maharah kalam menuntut instrumen yang mampu mengukur kemampuan berbicara secara valid dan reliabel, tetapi praktik di lapangan masih banyak bertumpu pada penilaian intuitif tanpa rubrik baku. Kajian ini bertujuan merumuskan tahapan pengembangan instrumen evaluasi maharah kalam berdasarkan kajian teoretis dan hasil penelitian terdahulu. Dengan metode kualitatif deskriptif melalui studi kepustakaan (library research), kajian ini menelaah teori evaluasi bahasa, model pengembangan instrumen (Mardapi, 4D, 3D, dan ADDIE), serta sepuluh hasil penelitian terdahulu yang terbit pada 2021-2026. Hasil kajian menunjukkan bahwa penelitian terdahulu umumnya melalui analisis kebutuhan, perancangan tugas dan rubrik, serta validasi ahli, tetapi bukti reliabilitas—terutama reliabilitas antarpenilai—jarang dilaporkan. Berdasarkan sintesis tersebut dirumuskan tujuh tahapan pengembangan: (1) analisis kebutuhan dan penetapan konstruk; (2) penyusunan spesifikasi dan kisi-kisi; (3) perancangan tugas dan rubrik; (4) validasi ahli; (5) uji coba terbatas; (6) analisis empiris; dan (7) revisi, perakitan, serta pelaporan. Rumusan ini diharapkan menjadi pegangan bagi pendidik dan peneliti dalam menyusun instrumen evaluasi maharah kalam yang sistematis dan dapat dipertanggungjawabkan.

Kata Kunci: maharah kalam; instrumen evaluasi; tahapan pengembangan; rubrik penilaian; kajian kepustakaan

Bagaimana Cara Sitasi Artikel ini:

Sulaiman, S., & Hamid, A. A. M. . (2026). Tahapan Pengembangan Instrumen Evaluasi Maharah Kalam Berdasarkan Kajian Teoretis dan Hasil Penelitian Terdahulu. Educational Journal, 2(2), 516-527. https://doi.org/10.63822/2cnbpb11

PENDAHULUAN

Maharah kalam menempati posisi sentral dalam pembelajaran bahasa Arab karena ia merupakan keterampilan produktif yang paling langsung memperlihatkan kemampuan seseorang menggunakan bahasa untuk berkomunikasi. Maharah kalam dipahami sebagai kemampuan mengungkapkan gagasan, pendapat, dan perasaan secara lisan kepada lawan bicara dengan pelafalan, kosakata, dan struktur yang tepat sehingga pesan dapat dipahami. Pandangan ini sejalan dengan konsep kompetensi komunikatif yang tidak membatasi kemampuan berbahasa pada penguasaan kaidah gramatikal, tetapi juga mencakup kompetensi sosiolinguistik, wacana, dan strategis.

Karena itu, evaluasi maharah kalam tidak cukup dilakukan dengan tes tertulis tentang kaidah. Evaluasi tersebut menuntut instrumen yang mampu menangkap performa lisan peserta didik secara valid dan reliabel. Penilaian berbicara memiliki tantangan yang khas: performanya bersifat sesaat, melibatkan banyak aspek sekaligus, dan dinilai oleh manusia sehingga rentan terhadap subjektivitas penilai.

Kenyataan di lapangan menunjukkan bahwa kebutuhan tersebut belum sepenuhnya terpenuhi. Penelitian Hasim dkk. di madrasah tsanawiyah pesantren di Riau mendapati bahwa banyak guru menilai kemampuan berbicara secara intuitif tanpa rubrik tertulis, karena tidak memiliki acuan yang pasti. Temuan senada dilaporkan Alamsyah dkk., yang menemukan bahwa penilaian bahasa Arab di sebuah madrasah aliyah cenderung bersifat administratif, berbasis hafalan, dan bertumpu pada tes tertulis, serta Zahroh dkk., yang mencatat tidak adanya rubrik baku, dokumentasi digital, dan penilaian berkelanjutan pada praktik penilaian kalam di sebuah program bahasa Arab.

Di sisi lain, beberapa tahun terakhir muncul sejumlah penelitian yang mengembangkan instrumen evaluasi kalam dengan beragam model dan tahapan, mulai dari model 4D, 3D, dan ADDIE hingga desain dan pengembangan berbasis CEFR. Namun tahapan yang ditempuh, bukti validitas dan reliabilitas yang dilaporkan, serta kedalaman prosedurnya berbeda-beda, sehingga belum tampak rumusan tahapan yang dapat dijadikan pegangan oleh pendidik maupun peneliti pemula.

Berdasarkan latar belakang tersebut, kajian ini merumuskan dua pertanyaan: (1) apa saja tahapan pengembangan instrumen evaluasi yang dikemukakan dalam kajian teoretis; dan (2) bagaimana tahapan tersebut diterapkan dalam penelitian-penelitian terdahulu, sehingga dapat dirumuskan tahapan pengembangan instrumen evaluasi maharah kalam yang sistematis? Kajian ini menggunakan metode kualitatif deskriptif dengan pendekatan studi kepustakaan (library research). Data dihimpun dari buku teori pengujian bahasa dan pengukuran pendidikan serta artikel jurnal terbitan 2015-2026 yang membahas pengembangan atau perancangan instrumen evaluasi maharah kalam, kemudian dianalisis melalui analisis isi (content analysis) dengan membandingkan model, tahapan, dan bukti kualitas instrumen pada setiap sumber.

LANDASAN TEORITIS

A. Maharah Kalam dan Evaluasinya

1. Hakikat Maharah Kalam dan Aspek yang Dinilai

Definisi operasional tentang apa yang diukur (konstruk) merupakan titik tolak setiap instrumen. Luoma menekankan bahwa penilaian berbicara harus berangkat dari pendefinisian yang jelas tentang kemampuan berbicara yang hendak diukur, baru kemudian dirumuskan tugas dan kriteria penilaiannya.

Dalam pembelajaran bahasa Arab, aspek yang lazim dinilai dalam maharah kalam adalah pelafalan, tata bahasa, kosakata, kelancaran, dan pemahaman. Sudrajat dkk. menyebut kelima aspek ini dan menyajikannya dalam skala penilaian berjenjang model Foreign Service Institute (FSI), dari tingkat 1 (sangat rendah) hingga tingkat 5 (sangat baik). Rumusan serupa dipakai Zahroh dkk. dalam rubrik analitik berskala 1-5, sementara Raswan dkk. merinci sepuluh aspek penilaian: isi, tampilan, kepercayaan diri, gestur, kelancaran, kemulusan, volume, makhraj, intonasi, dan tekanan. Perbedaan jumlah aspek ini menunjukkan bahwa konstruk kalam dapat didefinisikan lebih sempit atau lebih luas sesuai tujuan evaluasi dan jenjang peserta didik.

2. Jenis dan Teknik Evaluasi Maharah Kalam

Secara umum, instrumen evaluasi dibedakan menjadi instrumen tes dan nontes. Tes kalam pada hakikatnya adalah tes performa: peserta didik diminta menghasilkan ujaran yang kemudian dinilai berdasarkan kriteria tertentu. Mahmudi dan Nurbaha mengidentifikasi empat jenis tes kalam, yaitu tes pelafalan, tes struktur kalimat lisan, tes dengan stimulus visual, dan tes penyampaian ekspresi. Marzun menyebut bentuk tes lisan yang lebih tradisional, seperti membaca nyaring, setoran hafalan, tanya jawab langsung, pembacaan syair, dan diskusi laporan.

Dari sisi tingkat keterkendalian respons, Brown dan Abeywickrama membedakan tugas berbicara menjadi lima kategori: imitatif, intensif, responsif, interaktif, dan ekstensif, dari yang paling terkendali (mengulang) hingga yang paling bebas (monolog). Sejalan dengan itu, Sudrajat dkk. memetakan teknik tes berdasarkan jenjang: sepuluh teknik untuk tingkat pemula (misalnya mengulang, menyebutkan nama benda, dan melengkapi kalimat), sepuluh untuk tingkat menengah (misalnya mendeskripsikan, menceritakan gambar berseri, dan bermain peran), serta enam untuk tingkat lanjut (misalnya bercerita, wawancara, diskusi, dan pidato). Pemetaan ini menegaskan prinsip bahwa jenis tugas harus sepadan dengan jenjang kemampuan peserta didik.

Selain tes, penilaian nontes semakin mendapat tempat. Zahroh dkk. mengembangkan portofolio yang memadukan penilaian diri, penilaian sejawat, dan penilaian kinerja berbasis rekaman video, sedangkan Sabrina dan Baroroh memanfaatkan video dialog kreatif (hiwar) sebagai instrumen nontes kalam di madrasah aliyah. Tinjauan sistematis Clarisa dkk. atas literatur 2020-2025 juga menggambarkan pergeseran dari penilaian konvensional yang berfokus pada tata bahasa dan hafalan menuju model digital dan autentik yang menekankan kompetensi komunikatif.

3. Kriteria Instrumen yang Baik

Bachman dan Palmer memandang kegunaan (usefulness) sebuah tes sebagai hasil keseimbangan enam kualitas, yaitu reliabilitas, validitas konstruk, autentisitas, interaktivitas, dampak, dan praktikalitas; bobot masing-masing kualitas disesuaikan dengan konteks penggunaan tes. Kerangka ini penting karena instrumen kalam yang sangat autentik, misalnya wawancara bebas, sering kali sulit dinilai secara konsisten dan memakan waktu, sehingga pengembang harus mempertimbangkan kompromi secara sadar.

Validitas tidak dipahami sebagai sifat tes semata, melainkan sebagai penilaian evaluatif terpadu tentang sejauh mana bukti empiris dan rasional mendukung ketepatan interpretasi dan penggunaan skor. Untuk validitas isi, kesepakatan para ahli lazim dikuantifikasi, antara lain dengan koefisien V dari Aiken atau Content Validity Ratio (CVR) dari Lawshe. Adapun reliabilitas pada tes objektif diukur melalui

konsistensi internal, sedangkan pada performa lisan persoalan utamanya adalah konsistensi antarpenilai, yang dijaga melalui rubrik yang eksplisit dan pelatihan penilai.

B. Model dan Tahapan Pengembangan Instrumen: Kajian Teoretis

1. Tahapan Penyusunan Tes menurut Mardapi

   Mardapi merumuskan rangkaian langkah penyusunan tes yang secara garis besar meliputi penyusunan spesifikasi tes, penulisan soal, telaah soal, uji coba, analisis butir soal, perbaikan tes, perakitan tes, pelaksanaan tes, dan penafsiran hasil tes; sebagian peneliti memakai versi yang lebih ringkas dari rangkaian ini. Rangkaian tersebut menegaskan bahwa kualitas instrumen tidak ditentukan oleh penulisan butir semata, melainkan oleh siklus telaah, uji coba, analisis empiris, dan perbaikan.

   Meskipun dirumuskan terutama untuk tes tertulis, logika yang sama dapat diterapkan pada instrumen kalam dengan penyesuaian: "butir" berupa tugas lisan beserta rubriknya, dan analisis butir dilengkapi dengan analisis konsistensi penskoran antarpenilai.

2. **Model Pengembangan yang Dipakai dalam Penelitian**

   Dalam penelitian pengembangan, model 4D dari Thiagarajan dkk. dengan tahap define, design, develop, dan disseminate merupakan salah satu yang paling sering dirujuk. Raswan dkk. memakai model ini secara utuh untuk instrumen penilaian autentik kalam di madrasah aliyah, sedangkan Zahroh dkk. memakai versi tiga tahap (define, design, develop) tanpa tahap diseminasi. Gazali dan Saefuloh memakai ADDIE (analysis, design, development, implementation, evaluation) untuk tes kemahiran reseptif berbasis CEFR, Nurfadlilah dkk. memakai Design and Development Research (DDR) enam tahap untuk prototipe asesmen kalam CEFR B1, dan Hasibuan dkk. merujuk kerangka delapan langkah penyusunan tes dari Ainin.

   Meskipun nama dan jumlah tahapnya berbeda, model-model tersebut memiliki inti yang sama: (a) analisis kebutuhan, (b) perancangan, (c) pengembangan disertai validasi ahli, dan (d) uji coba serta evaluasi. Inti inilah yang menjadi kerangka untuk menelaah penelitian terdahulu pada bagian berikut.

3. **Teknik Pengujian Kualitas Instrumen**

   Pada penelitian terdahulu, validitas isi diuji melalui penilaian ahli yang dikuantifikasi dengan persentase kesepakatan, Aiken's V (ambang $V \geq 0{,}80$), atau Content Validity Index (CVI). Reliabilitas dilaporkan sebagai konsistensi internal, misalnya Cronbach's Alpha dengan ambang $\alpha \geq 0{,}70$ pada Zahroh dkk. dan Kuder-Richardson 20 (KR-20) pada Gazali dan Saefuloh. Untuk butir objektif, analisis tingkat kesukaran dan daya beda ditambahkan, seperti pada Syihabuddin dkk. dan Gazali dan Saefuloh.

C. Hasil Penelitian Terdahulu

1. Gambaran Prosedur dan Bukti Kualitas Instrumen

Tabel 1 merangkum sepuluh penelitian terdahulu yang ditelaah beserta model, tahapan, dan bukti validitas serta reliabilitas yang dilaporkan. Dua di antaranya (Gazali dan Saefuloh; Syihabuddin dkk.) berfokus pada tes kemahiran bahasa Arab secara umum, bukan khusus kalam, tetapi dimasukkan karena prosedur pengembangan dan analisis empirisnya relevan bagi instrumen kalam.

Tabel 1. Ringkasan penelitian terdahulu tentang pengembangan instrumen evaluasi maharah kalam dan tes kemahiran bahasa Arab

Peneliti (tahun)	Fokus dan model	Tahapan atau prosedur	Bukti validitas dan reliabilitas
Raswan dkk. (2022)	Penilaian autentik kalam MA; model 4D	Define (studi pustaka, wawancara, observasi), design (pemetaan KD, pedoman observasi), develop (revisi pakar), disseminate (uji terbatas)	Validitas pakar 84%, uji publik 78%; reliabilitas tidak dilaporkan
Zahroh dkk. (2026)	Portofolio autentik berbasis teknologi; model 3D	Define (analisis kebutuhan), design (lima indikator, skala 1-5), develop (validasi 5 ahli, uji terbatas 32 mahasiswa)	Aiken's V = 0,87; Cronbach's Alpha = 0,91
Nurfadlilah dkk. (2026)	Prototipe asesmen kalam CEFR B1; DDR	Enam tahap: identifikasi masalah, analisis deskriptor, blueprint, desain tugas, rubrik analitik, perakitan prototipe	Belum diuji secara empiris
Hasibuan dkk. (2024)	Butir tes menyimak dan kalam kelas XII MA; studi pustaka	Delapan langkah: tujuan, materi, topik, format, jumlah butir, kisi-kisi, penulisan, telaah	Tidak ada uji empiris
Alamsyah dkk. (2025)	Instrumen kosakata dalam berbicara, kelas X MA	Identifikasi masalah (observasi, wawancara, dokumen), rancangan tugas fungsional dan rubrik analitik	Respons guru positif; belum diuji kuantitatif
Sabrina dan Baroroh (2025)	Penilaian nontes: video dialog kreatif	Tugas video dialog dengan kriteria kelancaran, pelafalan, tata bahasa, keaktifan	Tidak dilaporkan
Gazali dan Saefuloh (2023)	Tes kemahiran reseptif berbasis CEFR; ADDIE	Analisis, desain (100 butir), pengembangan (8 ahli), implementasi (15 mahasiswa), evaluasi	S-CVI 0,96-0,97; KR-20 = 0,95
Syihabuddin dkk. (2021)	Tes kemahiran internet (Wikalah iBT)	Analisis kebutuhan, penulisan 140 butir, telaah kebahasaan, uji validitas dan reliabilitas, perancangan aplikasi, uji coba	81 butir valid; reliabilitas rata-rata 0,67
Hasim dkk. (2025)	Analisis kebutuhan instrumen kalam MTs pesantren	Angket, wawancara, observasi, dokumen	Tidak berlaku (analisis kebutuhan)
Clarisa dkk. (2026)	Tinjauan sistematis evaluasi kalam era digital	Studi pustaka terbitan 2020-2025	Tidak berlaku (tinjauan pustaka)

Penelitian yang melakukan validasi empiris paling lengkap ialah Raswan dkk. dan Zahroh dkk. Raswan dkk. menempuh studi pustaka, wawancara dengan pakar dan guru, serta observasi pada tahap define; memetakan kompetensi dasar dan menyusun pedoman observasi pada tahap design; merevisi draf berdasarkan masukan pakar pada tahap develop; lalu menguji secara terbatas pada tahap disseminate. Skala

awal 1-5 dipersempit menjadi 1-4 atas saran pakar demi kemudahan penggunaan guru, dan pakar juga mengkritik adanya indikator yang tumpang-tindih, misalnya intonasi dengan volume serta kelancaran dengan kemulusan. Validitas dilaporkan sebesar 84% pada uji pakar dan 78% pada uji publik, tetapi koefisien reliabilitas tidak disajikan. Zahroh dkk. menempuh analisis kebutuhan, perancangan portofolio dengan lima indikator berskala 1-5, validasi lima ahli, dan uji terbatas pada 32 mahasiswa, dengan hasil Aiken's V = 0,87 dan Cronbach's Alpha = 0,91. Namun, angka reliabilitas antarpenilai tidak dilaporkan, dan penulisnya mengakui keterbatasan berupa konteks tunggal, sampel kecil, serta fokus pada kelayakan, bukan efektivitas.

Dua penelitian pada ranah tes kemahiran memperlihatkan pentingnya analisis empiris. Gazali dan Saefuloh mengembangkan 100 butir pilihan ganda melalui lima tahap ADDIE, divalidasi oleh delapan ahli (S-CVI sangat tinggi, 0,96-0,97) dan diujicobakan pada 15 mahasiswa secara daring dengan KR-20 sebesar 0,95; sebanyak 97 butir dipakai tanpa revisi dan 3 butir direvisi. Sebaliknya, Syihabuddin dkk. yang menyusun 140 butir menemukan 25 butir bermasalah secara kebahasaan pada telaah awal, hanya 81 butir yang valid pada uji empiris, reliabilitas rata-rata 0,67 (terendah 0,50 pada bagian kaidah), dan sebaran tingkat kesukaran yang tidak proporsional. Kedua temuan ini menunjukkan bahwa butir yang tampak baik secara rasional belum tentu baik secara empiris.

Kelompok berikutnya berupa rancangan atau kajian konseptual yang belum diuji secara empiris. Nurfadlilah dkk. menyusun prototipe asesmen kalam CEFR B1 melalui enam tahap (identifikasi masalah, analisis deskriptor, pemetaan blueprint, perancangan tugas, penyusunan rubrik analitik, dan perakitan prototipe) yang menghasilkan 15 tugas dengan komposisi 3 tugas pemahaman, 8 tugas interaksi, dan 4 tugas produksi; penulisnya sendiri menyatakan bahwa validasi ahli, uji coba, dan analisis psikometrik masih perlu dilakukan. Hasibuan dkk. menerapkan delapan langkah penyusunan tes pada tema olahraga untuk kelas XII madrasah aliyah dan menghasilkan kisi-kisi tiga tugas kalam, tetapi tanpa uji empiris. Alamsyah dkk. mengidentifikasi masalah melalui observasi, wawancara, dan telaah dokumen, lalu merancang tugas fungsional (deskripsi gambar, bercerita, dan bermain peran) dengan rubrik analitik yang menilai ketepatan makna, kontrol gramatikal, dan ketepatan pragmatik; respons guru positif, tetapi efektivitasnya belum diuji secara kuantitatif. Sabrina dan Baroroh mengembangkan instrumen nontes berbasis video dialog dengan kriteria kelancaran, pelafalan, tata bahasa, dan keaktifan.

Terakhir, dua penelitian memberi landasan kebutuhan. Hasim dkk. melakukan analisis kebutuhan melalui angket, wawancara, observasi, dan dokumen, serta menemukan bahwa sekitar separuh siswa menilai petunjuk tes kurang jelas dan seluruh guru sepakat bahwa pelatihan penyusunan dan penilaian sangat diperlukan. Clarisa dkk. menegaskan peluang sekaligus tantangan penilaian kalam era digital, yakni literasi digital guru, infrastruktur, dan kekhawatiran terhadap validitas serta objektivitas.

2. Pola, Kekuatan, dan Celah

Pertama, analisis kebutuhan hampir selalu menjadi tahap awal, baik melalui observasi, wawancara, maupun telaah dokumen dan kurikulum (Raswan dkk., Zahroh dkk., Gazali dan Saefuloh, Alamsyah dkk., Hasim dkk., dan Syihabuddin dkk.). Kedua, aspek yang dinilai cenderung konvergen pada pelafalan, kelancaran, kosakata, tata bahasa, dan pemahaman atau isi, tetapi perumusan indikatornya masih rawan tumpang-tindih sebagaimana dikritik pakar pada penelitian Raswan dkk. Ketiga, validasi ahli menjadi bukti validitas isi yang seragam dengan jumlah ahli empat sampai delapan orang, namun cara mengkuantifikasinya berbeda-beda (persentase, Aiken's V, atau CVI) sehingga hasil antarpenelitian sulit dibandingkan.

Keempat, dan ini celah terpenting, bukti reliabilitas masih lemah. Dari sepuluh penelitian yang ditelaah, hanya tiga yang melaporkan koefisien reliabilitas (Zahroh dkk., Gazali dan Saefuloh, serta Syihabuddin dkk.), dan satu-satunya yang berkaitan langsung dengan kalam, yaitu Zahroh dkk., melaporkan konsistensi internal tanpa angka reliabilitas antarpenilai. Padahal subjektivitas penilai merupakan risiko utama penilaian berbicara. Kelima, tren penilaian bergerak ke arah autentik dan digital (portofolio video, video dialog, dan platform daring), tetapi pengembangan ini menuntut perhatian pada kesiapan guru dan infrastruktur. Keenam, sebagian besar penelitian dilakukan pada konteks tunggal dengan sampel kecil sehingga generalisasinya terbatas. Celah-celah inilah yang menjadi dasar perumusan tahapan terpadu pada bagian berikut.

D. Rumusan Tahapan Pengembangan Instrumen Evaluasi Maharah Kalam

Berdasarkan kajian teoretis pada bagian C dan pola penelitian terdahulu pada bagian D, penulis merumuskan tujuh tahapan pengembangan instrumen evaluasi maharah kalam sebagaimana diringkas pada Tabel 2. Rumusan ini merupakan sintesis penulis dari sumber-sumber yang dikaji, bukan model baku yang telah diuji secara empiris, dan urutannya bersifat iteratif: hasil pada tahap validasi, uji coba, atau analisis dapat mengembalikan pengembang ke tahap perancangan.

Tabel 2. Tujuh tahapan pengembangan instrumen evaluasi maharah kalam (sintesis penulis)

No.	Tahap	Kegiatan inti	Keluaran
1	Analisis kebutuhan dan penetapan konstruk	Observasi, wawancara guru, telaah kurikulum dan praktik penilaian; tetapkan tujuan, jenjang, dan konstruk	Laporan kebutuhan; definisi konstruk
2	Spesifikasi dan kisi-kisi	Tetapkan tema, indikator, jenis dan jumlah tugas, format respons, durasi, dan bobot	Cetak biru instrumen
3	Perancangan tugas dan rubrik	Tulis tugas autentik; susun rubrik analitik dan pedoman penskoran	Draf tugas, rubrik, dan pedoman
4	Validasi ahli	Telaah materi, konstruk, bahasa, dan keterbacaan; kuantifikasi dengan Aiken's V atau CVI	Instrumen tervalidasi dan direvisi
5	Uji coba terbatas	Uji pada sampel kecil; nilai kejelasan petunjuk, waktu, dan kelayakan administrasi	Catatan uji coba dan revisi
6	Analisis empiris	Hitung reliabilitas antarpenilai dan konsistensi internal; analisis butir bila ada butir objektif	Bukti reliabilitas dan kualitas butir
7	Revisi, perakitan, dan pelaporan	Finalisasi paket instrumen; pelatihan penilai; pelaporan dan evaluasi berkala	Instrumen final dan panduan penggunaan

1. Analisis Kebutuhan dan Penetapan Konstruk

Tahap pertama bertujuan memastikan bahwa instrumen yang dikembangkan memang dibutuhkan dan sesuai dengan konteks. Kegiatannya meliputi observasi dan wawancara dengan guru, telaah kurikulum, capaian pembelajaran, dan buku ajar, serta analisis praktik penilaian yang berlaku. Tahap ini juga menetapkan tujuan evaluasi (diagnostik, formatif, atau sumatif), karakteristik dan jenjang peserta didik, serta konstruk kalam yang akan diukur. Tahap serupa ditempuh pada fase define oleh Raswan dkk. dan Zahroh dkk. serta pada langkah awal penelitian Alamsyah dkk., dan menjadi langkah pertama dalam kerangka delapan langkah yang dipakai Hasibuan dkk. (menentukan tujuan penilaian). Rumusan konstruk

dapat merujuk pada kompetensi komunikatif dan acuan jenjang seperti FSI atau CEFR agar tingkat kemampuan yang diukur jelas.

2. Penyusunan Spesifikasi dan Kisi-Kisi

Tahap kedua menerjemahkan konstruk menjadi cetak biru (blueprint) instrumen: tema atau materi, indikator, jenis dan jumlah tugas, format respons, durasi, serta bobot setiap aspek. Pada tahap ini jenis tugas dipilih sepadan dengan jenjang, misalnya tugas imitatif dan intensif untuk pemula serta tugas interaktif dan ekstensif untuk tingkat lanjut. Nurfadlilah dkk. memperlihatkan praktik pemetaan deskriptor kemampuan ke dalam indikator terukur dan mendistribusikan tugas menurut jenisnya, sedangkan Hasibuan dkk. menyusun kisi-kisi yang memuat tema, format, dan jumlah butir sebelum menulis tugas. Langkah ini sesuai dengan tahap penyusunan spesifikasi tes dalam rangkaian langkah Mardapi.

3. Perancangan Tugas dan Rubrik Penilaian

Tahap ketiga menghasilkan tugas kalam beserta rubrik dan pedoman penskorannya. Tugas sebaiknya autentik dan bermakna, misalnya memperkenalkan diri, bermain peran, mendeskripsikan gambar, menyampaikan pendapat, bercerita, dan wawancara, sebagaimana delapan jenis tugas yang diidentifikasi Zahroh dkk. Rubrik analitik lebih dianjurkan karena memberi umpan balik per aspek; setiap aspek disertai deskriptor yang jelas untuk setiap tingkat skor. Pelajaran dari Raswan dkk. perlu diperhatikan: indikator yang tumpang-tindih harus dihindari dan jumlah tingkat skala dipilih dengan mempertimbangkan kemudahan penggunaan oleh guru. Bila relevan, pengembang dapat melengkapi penilaian guru dengan penilaian diri dan sejawat serta pengumpulan performa melalui rekaman video.

4. Validasi Ahli

Tahap keempat menguji validitas isi dan keterbacaan instrumen oleh para ahli, yaitu ahli bahasa Arab, ahli evaluasi atau pengukuran, dan praktisi (guru). Jumlah ahli pada penelitian terdahulu berkisar antara empat dan delapan orang. Aspek yang ditelaah meliputi kesesuaian dengan konstruk dan indikator, kejelasan tugas dan rubrik, ketepatan bahasa, serta kelayakan penerapan. Kesepakatan ahli sebaiknya dikuantifikasi dengan Aiken's V (ambang $V \geq 0{,}80$) atau CVR/CVI, dan dilengkapi telaah kebahasaan karena Syihabuddin dkk. menemukan kesalahan pada 25 dari 140 butir pada tahap ini. Seluruh masukan ahli ditindaklanjuti dengan revisi.

5. Uji Coba Terbatas

Tahap kelima menguji kelayakan instrumen pada sampel kecil yang menyerupai sasaran sebenarnya. Fokusnya adalah kejelasan petunjuk, kecukupan waktu, kelancaran administrasi (termasuk perekaman bila memakai video), serta respons guru dan peserta didik. Hasim dkk. mengingatkan bahwa petunjuk yang kurang jelas dan waktu yang terlalu singkat merupakan keluhan nyata peserta didik, sedangkan Zahroh dkk. dan Gazali dan Saefuloh masing-masing memakai uji terbatas pada 32 dan 15 peserta. Bila memakai platform digital, jalur alternatif perlu disediakan bagi peserta didik yang terkendala akses.

6. Analisis Empiris: Reliabilitas dan Kualitas Butir

Tahap keenam menjawab celah yang paling menonjol pada penelitian terdahulu. Pada instrumen kalam, minimal dua penilai independen menilai rekaman atau performa yang sama menggunakan rubrik, kemudian tingkat kesepakatan mereka dihitung dengan koefisien yang sesuai (misalnya korelasi antarpenilai atau kappa). Konsistensi internal antaraspek dapat diuji dengan Cronbach's Alpha (ambang $\alpha \geq 0{,}70$), dan bila instrumen memuat butir objektif, analisis tingkat kesukaran dan daya beda dilakukan seperti pada Syihabuddin dkk. dan Gazali dan Saefuloh. Penentuan koefisien kesepakatan merupakan rekomendasi penulis, karena belum ada penelitian kalam yang ditelaah yang melaporkannya. Pentingnya

tahap ini terlihat dari temuan Syihabuddin dkk. bahwa hanya 81 dari 140 butir yang valid secara empiris. Langkah ini sejalan dengan analisis butir dan perbaikan tes dalam rangkaian langkah Mardapi.

7. Revisi, Perakitan, dan Pelaporan

Tahap terakhir merevisi instrumen berdasarkan hasil validasi, uji coba, dan analisis empiris, lalu merakitnya menjadi paket yang siap pakai: tugas, rubrik, pedoman penskoran, panduan administrasi, dan petunjuk bagi penilai. Pelatihan penilai perlu dipertimbangkan, mengingat seluruh guru dalam penelitian Hasim dkk. menyatakan pelatihan sebagai kebutuhan. Hasil pengembangan dilaporkan secara transparan, termasuk keterbatasannya, dan instrumen dievaluasi secara berkala karena kebutuhan serta teknologi penilaian terus berkembang. Tahap ini berpadu dengan langkah perakitan, pelaksanaan, dan penafsiran hasil tes pada rangkaian langkah Mardapi.

Secara keseluruhan, ketujuh tahapan tersebut tidak harus dijalankan sama dalamnya. Untuk evaluasi kelas yang bersifat formatif, guru dapat menyederhanakan tahap uji coba dan analisis empiris; untuk ujian sumatif atau tes berisiko tinggi, tahap validasi dan analisis reliabilitas tidak boleh dilewati. Prinsip yang perlu dijaga adalah keseimbangan antara validitas, reliabilitas, autentisitas, dan praktikalitas.

KESIMPULAN

Berdasarkan kajian di atas, dapat disimpulkan bahwa pengembangan instrumen evaluasi maharah kalam yang baik memerlukan prosedur yang sistematis, bukan sekadar penyusunan soal atau rubrik. Kajian teoretis (tahapan penyusunan tes Mardapi serta model 4D, 3D, ADDIE, dan DDR) dan penelitian terdahulu memiliki inti yang sama, yaitu analisis kebutuhan, perancangan, validasi ahli, serta uji coba dan evaluasi. Dari sintesis tersebut dirumuskan tujuh tahapan: analisis kebutuhan dan penetapan konstruk; penyusunan spesifikasi dan kisi-kisi; perancangan tugas dan rubrik; validasi ahli; uji coba terbatas; analisis empiris; serta revisi, perakitan, dan pelaporan.

Penelitian terdahulu relatif kuat pada tahap analisis kebutuhan dan validasi ahli, tetapi lemah pada bukti reliabilitas, khususnya reliabilitas antarpenilai pada penilaian performa lisan. Kajian ini bersifat konseptual dan berbasis kepustakaan, sehingga rumusan tahapan yang diajukan masih perlu diuji melalui penelitian pengembangan empiris, dengan sampel yang lebih besar, konteks yang lebih beragam, dan pelaporan reliabilitas antarpenilai yang eksplisit. Bagi pendidik, rumusan ini dapat dipakai sebagai daftar periksa untuk menyusun atau menelaah instrumen kalam yang digunakan di kelasnya.

DAFTAR PUSTAKA

Aiken, Lewis R. "Three Coefficients for Analyzing the Reliability and Validity of Ratings." Educational and Psychological Measurement 45, no. 1 (1985): 131-142.

Alamsyah, Firman, Rival Fitroh, dan R. Umi Baroroh. "Innovative Arabic Vocabulary Assessment Instruments to Enhance the Speaking Skills of Grade X Students at MA Science and Technology Al Marzuqiyyah Sumedang." AL-I'ROBY: Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban 2, no. 2 (2025): 82-96.

Bachman, Lyle F., dan Adrian S. Palmer. Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press, 1996.

Brown, H. Douglas, dan Priyanvada Abeywickrama. Language Assessment: Principles and Classroom Practices. Ed. ke-2. White Plains, NY: Pearson Education, 2010.

Canale, Michael, dan Merrill Swain. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing." Applied Linguistics 1, no. 1 (1980): 1-47.

Clarisa, Vivi Dwi, dkk. "Problematika dan Transformasi Evaluasi Mahārah Kalām: Tinjauan Sistematis dalam Pembelajaran Bahasa Arab Era Digital." Al Ibrah: Journal of Arabic Language Education 9, no. 1 (2026): 145-151.

Effendy, Ahmad Fuad. Metodologi Pengajaran Bahasa Arab. Malang: Misykat, 2017.

Fulcher, Glenn. Testing Second Language Speaking. Harlow: Pearson Longman, 2003.

Gazali, Erfan, dan Hasan Saefuloh. "Development of an Arabic Receptive Proficiency Test Instrument Based on the Common European Framework of Reference for Languages." Al-Ta'rib: Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab 11, no. 2 (2023): 293-308.

Hasibuan, Renni, Dewi Chamidah, dan Akhsan. "Tathbīq I'dād Bunūd al-Ikhtibār li Mahārah al-Istimā' wa al-Kalām." Lahjah Arabiyah: Jurnal Bahasa Arab dan Pendidikan Bahasa Arab 5, no. 1 (2024): 15-30.

Hasim, Muhamad Fuad, dkk. "Bridging Assessment and Proficiency: Developing Speaking Skill Test Instruments for Arabic Learners." Tsaqofiya: Jurnal Pendidikan Bahasa dan Sastra Arab 7, no. 2 (2025): 526-537.

Lawshe, C.H. "A Quantitative Approach to Content Validity." Personnel Psychology 28, no. 4 (1975): 563-575.

Luoma, Sari. Assessing Speaking. Cambridge: Cambridge University Press, 2004.

Mahmudi, Ihwan, dan Ayu Nurbaha. "Anwā' al-Ikhtibārāt wa al-Mustawayāt al-Mi'yāriyyah li Mahārah al-Kalām fī al-Lugah al-'Arabiyyah." Aphorisme: Journal of Arabic Language, Literature, and Education 4, no. 2 (2023).

Mardapi, Djemari. Teknik Penyusunan Instrumen Tes dan Nontes. Yogyakarta: Mitra Cendikia Press, 2008.

Marzun R. "al-Ikhtibār al-Syafahī wa Isti'māluhu fī Taqwīm Mahārah al-Kalām." Lisanuna: Jurnal Ilmu Bahasa Arab dan Pembelajarannya 4, no. 2 (2015): 202-218.

Messick, Samuel. "Validity." Dalam Educational Measurement, disunting oleh Robert L. Linn, ed. ke-3, 13-103. New York: American Council on Education dan Macmillan, 1989.

Nurfadlilah, Siti, dkk. "Designing a CEFR B1-Based Arabic Speaking Assessment Prototype through the Adaptation of Ali Al-Khuli's Speaking Task Framework." Ta'lim al-'Arabiyyah: Jurnal Pendidikan Bahasa Arab & Kebahasaaraban 10, no. 1 (2026): 36-53.

Raswan, N. Lalah Alawiyah, dan Taufik Luthfi. "Model of Arabic Authentic Assessment Instruments: Speaking (Kalām) at Madrasah Aliyah." Alsinatuna: Journal of Arabic Linguistics and Education 8, no. 1 (2022): 51-64.

Sabrina, Banat, dan R. Umi Baroroh. "Inovasi Instrumen Penilaian Non Tes Maharah Kalam dengan Pembuatan Video Hiwar Kreatif di MAN 4 Banyuwangi." Pendas: Jurnal Ilmiah Pendidikan Dasar 10, no. 2 (2025): 446-459.

Sudrajat, Azzahra Emira, dkk. "Instrumen Evaluasi Maharah Kalam dalam Pembelajaran Bahasa Arab." Al-Tadris: Jurnal Pendidikan Bahasa Arab 13, no. 2 (2025): 615-641.

Syihabuddin, dkk. "Wikalah iBT: How Is an Internet-Based Arabic Language Proficiency Test Instrument Developed?" Ijaz Arabi: Journal of Arabic Learning 4, no. 3 (2021): 535-547.

Thiagarajan, Sivasailam, Dorothy S. Semmel, dan Melvyn I. Semmel. Instructional Development for Training Teachers of Exceptional Children: A Sourcebook. Bloomington: Center for Innovation in Teaching the Handicapped, Indiana University, 1974.

Zahroh, Fajrotuz, dkk. "Developing a Technology-Integrated Authentic Pragmatic Portfolio Assessment for Mahārah al-Kalām in the Malzamah Akbarnas Program." Alsina: Journal of Arabic Studies 8, no. 1 (2026): 97-130.