



Asesmen HOTS Bahasa Arab Berbasis Bukti Penalaran dalam Lingkungan Digital

Moh. Ainin, Moch Wahib Dariyadi

Universitas Negeri Malang, Indonesia

Email: moh.ainin.fs@um.ac.id; wahib.fs@um.ac.id — Korespondensi: wahib.fs@um.ac.id

Abstract

Higher-order Arabic assessment should elicit defensible reasoning rather than infer cognitive complexity from long texts or difficult vocabulary. This article develops an evidence-oriented framework for digital assessment through a purposive narrative review and analysis of constructed tasks. Selected Arabic assessment studies are connected with revised Bloom categories, argument-based validation, and formative assessment. The synthesis distinguishes the intended claim, the task that elicits evidence, the learner response, and the interpretation of the resulting score. A short Arabic announcement illustrates how the same stimulus can support factual retrieval, inference, evaluation, or text revision depending on the question and scoring criteria. Digital delivery is positioned as a way to preserve explanations and revisions, not as a guarantee of higher-order thinking. The framework proposes an assessment blueprint, a two-stage response format, and an initial analytic rubric separating answer accuracy, textual evidence, and inferential justification. It also identifies vocabulary burden, ambiguous alternatives, technological barriers, and feedback timing as potential threats to score interpretation. No learner data, item statistics, or validity coefficients are reported. The contribution is an untested design framework that requires expert review, response-process investigation, independent scoring, and classroom piloting before it can support decisions about learners' higher-order Arabic abilities.

Keywords	Arabic assessment; higher-order thinking; reasoning evidence; digital assessment; score interpretation
----------	--

مستخلص البحث

ينبغي أن يستثير تقييم التفكير العالي في العربية تعليلاً يمكن فحصه، لا أن يستدل على صعوبة التفكير من طول النص أو غرابة المفردات. تهدف المقالة إلى بناء إطار للتقويم الرقمي القائم على الأدلة من خلال مراجعة سردية قصدية وتحليل مهام مصنوعة. وترتبط دراسات تقويم العربية بتصنيف بلوم المعدل والتحقق القائم على الحجة والتقويم التكويني. ويميز الإطار بين الادعاء المقصود والمهمة واستجابة المتعلم وتفسير الدرجة. ويوضح إعلان عربي قصير كيف يدعم النص نفسه استرجاع المعلومات أو الاستنتاج أو التقويم أو المراجعة بحسب السؤال ومعياري التصحيح. ويستخدم الوسيط الرقمي لحفظ التعليقات والمراجعات، ولا يعد ضماناً للتفكير العالي. وتقترح المقالة خريطة اختبار وصيغة جواب على مرحلتين وسلم تقدير يفصل صحة الجواب عن الدليل والتعليل. وتناقش عبء المفردات وغموض البدائل والحوازر التقنية وتوقيت التغذية الراجعة بوصفها تهديدات محتملة للتفسير. ولا تعرض بيانات طلاب أو إحصاءات فقرات أو معاملات صدق، لأن الإطار مقترح يحتاج إلى تحكيم ودراسة عمليات الاستجابة وتجربة صافية قبل استخدامه لاتخاذ قرارات تقويمية.

كلمات أساسية	تقويم العربية؛ التفكير العالي؛ أدلة التعليل؛ التقويم الرقمي؛ تفسير الدرجات
--------------	--

1 Pendahuluan

1.1 Kompleksitas berpikir dan kesulitan bahasa

Soal bahasa Arab dapat terasa sulit karena kosakata, panjang kalimat, atau ketidakakraban tema. Kesulitan tersebut tidak selalu berarti tuntutan berpikir tingkat tinggi. Sebaliknya, teks

singkat dengan kosakata sederhana dapat menuntut evaluasi yang kuat jika peserta harus membandingkan klaim dan menunjukkan bukti. Persoalan asesmen ialah bagaimana memisahkan kerumitan bahasa dari proses yang hendak diukur. Artikel ini menempatkan bukti penalaran sebagai penghubung tujuan, tugas, serta interpretasi hasil.

Penggunaan kata kerja “analisis” dalam perintah tidak cukup untuk menentukan kategori soal. Jika peserta hanya menyalin informasi yang disebut langsung, tugas masih dapat diselesaikan tanpa analisis hubungan. Demikian pula, meminta peserta “menciptakan” kalimat dengan meniru satu pola belum tentu menuntut penciptaan dalam pengertian yang luas. Kategori kognitif perlu diperiksa melalui tindakan yang diperlukan untuk menjawab dan bukan melalui pilihan istilah pada kisi-kisi. Prinsip ini penting bagi tugas cetak maupun digital.

Zubaidah et al. (2022) menganalisis tingkat instruksi pada buku bahasa Arab sekolah menengah. Masrifah et al. (2025) meneliti latihan HOTS dalam pembelajaran bahasa Arab melalui perbandingan kelas. Kedua sumber menyediakan konteks penting, tetapi kategori instruksi dan hasil intervensi merupakan jenis bukti berbeda. Artikel ini tidak memindahkan hasilnya menjadi klaim atas semua soal atau seluruh siswa. Fokusnya adalah menyusun rancangan yang membuat alasan pemberian kategori dan skor dapat ditelusuri.

1.2 Asesmen digital dan validitas

Lingkungan digital menyediakan ruang untuk menyimpan pilihan, uraian, waktu respons, dan revisi. Namun, banyaknya informasi yang direkam tidak menunjukkan bahwa informasi itu relevan dengan kemampuan. Waktu pengerjaan dapat dipengaruhi perangkat; jumlah klik dapat menunjukkan navigasi; perubahan jawaban dapat muncul karena menerima petunjuk. Interpretasi perlu membedakan proses penalaran dari perilaku penggunaan aplikasi. Teknologi membantu pengumpulan bukti hanya apabila hubungan bukti dengan klaim dinyatakan secara jelas.

Kane (2013) menempatkan validasi pada argumen mengenai interpretasi dan penggunaan skor. Orientasi ini berarti bahwa satu instrumen tidak cukup dinyatakan “valid” tanpa menyebut penggunaan yang dimaksud. Nilai dapat membantu merancang umpan balik tetapi belum memadai untuk keputusan yang lebih luas. Black dan Wiliam (1998) membahas asesmen kelas serta fungsi formatifnya. Dalam kerangka artikel ini, bukti dikumpulkan untuk memahami penalaran dan menentukan bantuan, sementara keputusan sumatif membutuhkan pengendalian tambahan.

1.3 Tujuan kajian

Artikel bertujuan merumuskan asesmen HOTS bahasa Arab yang menghubungkan klaim kemampuan dengan bukti respons. Pertanyaannya meliputi cara membedakan tuntutan kognitif, bentuk respons yang menunjukkan penalaran, dan ancaman interpretasi pada penyajian digital. Kontribusi berupa kerangka, contoh tugas buatan, kisi-kisi, serta rubrik awal. Tidak ada pengumpulan respons peserta atau analisis butir yang dilakukan. Klaim efektivitas dan mutu psikometrik tetap menjadi agenda pengujian berikutnya.

2 Metode

2.1 Pendekatan dan sumber

Kajian menggunakan tinjauan naratif purposif dan analisis konseptual tugas. Sumber inti membahas asesmen bahasa Arab, kategori tujuan pembelajaran, validasi interpretasi skor,

serta asesmen formatif. Penelusuran dilakukan melalui judul dan tema pada laman jurnal serta institusi. Pemilihan sumber tidak mengikuti prosedur tinjauan sistematis, sehingga tidak dilaporkan jumlah hasil pencarian atau diagram seleksi. Literatur dipakai untuk membangun hubungan antarkonsep, bukan menghitung prevalensi kualitas soal di lapangan.

Studi buku ajar dipakai untuk membahas tuntutan instruksi; studi kelas dipakai sebagai konteks penelitian HOTS; teori validasi dipakai untuk memeriksa batas klaim. Informasi yang dirujuk dibatasi pada bagian sumber yang dapat diperiksa. Perbedaan desain sumber dipertahankan dalam sintesis. Artikel tidak menggabungkan kategori instruksi dengan ukuran efek seolah-olah keduanya memberikan bukti yang sama tentang kemampuan individu.

2.2 Prosedur analisis

Analisis dilakukan dengan menuliskan klaim kemampuan, menentukan tindakan untuk memperoleh bukti, merancang stimulus, serta membandingkan respons yang mungkin muncul. Contoh pengumuman dibuat untuk memperlihatkan perubahan tuntutan pada stimulus yang sama. Contoh tidak berasal dari buku ajar yang diteliti, dokumen sekolah, atau hasil pekerjaan peserta. Sesudah itu, dirumuskan ancaman interpretasi dan cara mengendalikannya. Seluruh hasil berupa argumen desain yang dapat ditelaah, bukan temuan psikometrik.

Tidak dilakukan uji keterbacaan, validasi pakar independen, atau penghitungan reliabilitas. Rubrik dan kisi-kisi memiliki status usulan. Pemeriksaan internal hanya memastikan konsistensi antara teks, pertanyaan, dan alasan jawaban. Tahap ini berbeda dari validasi dengan data peserta. Peneliti berikutnya perlu memeriksa bagaimana peserta benar-benar mengerjakan tugas karena tindakan yang diharapkan penulis belum tentu sama dengan proses respons aktual.

3 Hasil Analisis Konseptual

3.1 Rantai klaim tugas bukti dan interpretasi

Kerangka menghasilkan empat komponen: klaim, tugas, bukti, dan interpretasi. Klaim menjelaskan kemampuan yang hendak diketahui, misalnya menilai kecukupan informasi sebuah pengumuman. Tugas meminta tindakan yang sesuai. Bukti berupa keputusan disertai alasan atau revisi. Interpretasi menghubungkan kualitas respons dengan kemampuan dalam batas tugas. Jika salah satu hubungan tidak jelas, penilai dapat menarik simpulan terlalu luas dari jawaban yang tampak benar.

Tabel 1 Rantai argumentasi asesmen

Komponen	Pertanyaan rancangan	Contoh penerapan
Klaim	Kemampuan apa yang hendak dijelaskan	Menilai kelengkapan informasi pengumuman
Tugas	Tindakan apa yang memunculkan bukti	Memilih informasi tambahan dan memberi alasan
Bukti	Respons apa yang relevan	Kutipan petunjuk serta alasan kebutuhan pembaca

Komponen	Pertanyaan rancangan	Contoh penerapan
Interpretasi	Simpulan sejauh apa yang diizinkan	Kualitas penalaran pada jenis pengumuman ini

Sumber: hasil analisis dan rancangan dalam artikel ini.

Klaim harus cukup spesifik. “Mampu berpikir kritis dalam bahasa Arab” terlalu luas jika hanya didasarkan pada satu soal. “Mampu membandingkan dua kemungkinan berdasarkan petunjuk teks” lebih dekat dengan bukti yang tersedia. Pembatasan tidak mengurangi nilai asesmen; ia membuat hasil dapat dipakai dengan tepat. Skor yang rendah kemudian diarahkan pada jenis kesulitan yang terlihat, bukan langsung pada label kemampuan umum.

3.2 Stimulus sama dan tuntutan berbeda

Stimulus buatan berbunyi «عُلِّقَ الْمَكْتَبَةُ يَوْمَ الْجُمُعَةِ بِسَبَبِ الصِّيَاةِ، وَتُفْتَحُ يَوْمَ السَّبْتِ فِي السَّاعَةِ الثَّامِنَةِ صَبَاحًا». Pengumuman menyatakan perpustakaan ditutup pada Jumat karena pemeliharaan dan dibuka pada Sabtu pukul delapan pagi. Pertanyaan tentang hari penutupan meminta informasi langsung. Pertanyaan tentang waktu kunjungan yang sesuai meminta pemanfaatan informasi, tetapi belum otomatis menjadi evaluasi. Jenis tindakan harus dilihat dari hubungan yang diperlukan untuk menjawab.

Untuk menuntut penalaran lebih kuat, tugas dapat menambahkan situasi siswa yang ingin mengembalikan buku pada Jumat. Peserta diminta menilai apakah pengumuman sudah cukup memberi solusi dan menjelaskan informasi yang masih dibutuhkan. Jawaban dapat menyebut cara pengembalian saat tutup atau batas keterlambatan, jika dijelaskan sebagai informasi yang tidak tersedia. Peserta tidak boleh mengarang bahwa fasilitas tersebut pasti ada. Kemampuan membedakan yang disebut dan yang belum diketahui menjadi bukti penting.

Tugas revisi meminta peserta memperbaiki pengumuman untuk pembaca yang memiliki kebutuhan tertentu. Ia dapat menambahkan kalimat contoh, tetapi harus menandainya sebagai usulan, bukan isi dokumen sumber. Penilai memeriksa konsistensi dengan informasi awal serta tujuan komunikasi. Dengan demikian, tugas penciptaan tidak hanya menghasilkan kalimat baru, melainkan menghubungkan revisi dengan masalah yang telah diidentifikasi. Teks pendek menyediakan ruang penalaran tanpa menambah kosakata yang tidak relevan.

Tabel 2 Perbedaan tuntutan pada stimulus buatan

Tindakan	Pertanyaan	Bukti yang membedakan
Mengambil fakta	Kapan perpustakaan tutup	Informasi langsung dari teks
Menggunakan informasi	Kapan kunjungan dapat dilakukan	Pilihan waktu yang konsisten
Menilai kecukupan	Apakah pengumuman menjawab kebutuhan pengembalian Jumat	Penjelasan informasi yang belum tersedia
Merevisi	Tambahkan informasi sesuai kebutuhan pembaca	Usulan dengan alasan dan batas sumber

Sumber: hasil analisis dan rancangan dalam artikel ini.

3.3 Respons dua tahap

Format dua tahap meminta peserta memberikan jawaban dan bukti alasan. Pada pilihan ganda, peserta memilih alternatif kemudian menyebut petunjuk pendukung. Pada uraian, peserta menuliskan keputusan singkat lalu penjelasannya. Format ini membantu membedakan tebakan, pemahaman parsial, dan argumen yang konsisten. Akan tetapi, respons alasan yang baik belum otomatis menunjukkan penalaran mandiri apabila peserta telah menerima model sebelum menjawab. Waktu pemberian petunjuk perlu dicatat.

Alasan tidak perlu panjang. Satu atau dua kalimat yang menghubungkan petunjuk dengan keputusan dapat lebih kuat daripada uraian yang mengulang seluruh stimulus. Penilai harus menerima variasi redaksi yang tepat. Jika tujuan utama ialah membaca, jawaban bahasa Indonesia dapat dipertimbangkan agar produksi Arab tidak mencampuri konstruk, tetapi keputusan itu dinyatakan dalam spesifikasi. Jika tujuan mencakup argumentasi Arab, aspek produksi bahasa dinilai terpisah dari ketepatan penalaran.

3.4 Pengecoh dan ambiguitas

Alternatif pengecoh perlu merepresentasikan kekeliruan yang masuk akal, bukan kalimat yang jelas tidak berhubungan. Pengecoh dapat mencerminkan salah membaca waktu, menganggap informasi tambahan tersedia, atau memilih solusi yang tidak konsisten. Namun, penilai tidak boleh merancang dua alternatif sama-sama benar tanpa kriteria pemilihan. Kualitas pengecoh diperiksa melalui alasan respons aktual, bukan diasumsikan dari niat menyusun soal. Artikel ini tidak mempunyai data untuk menentukan daya pengecoh.

Tabel 3 Ancaman dan pengendalian awal

Ancaman	Dampak pada interpretasi	Pengendalian
Kosakata berlebihan	Skor lebih mencerminkan hambatan leksikal	Glosarium untuk kata di luar target
Alternatif ambigu	Jawaban berbeda dapat sama-sama masuk akal	Kriteria dan telaah ahli
Petunjuk terlalu awal	Respons meniru model	Pisahkan respons awal dan revisi
Hambatan perangkat	Waktu tidak mencerminkan kemampuan bahasa	Uji tampilan dan akses
Skor gabungan	Penalaran tercampur ketepatan produksi	Pisahkan dimensi rubrik

Sumber: hasil analisis dan rancangan dalam artikel ini.

4 Pembahasan

4.1 Kisi kisi sebagai spesifikasi bukti

Kisi-kisi sebaiknya memuat tujuan, jenis teks, tindakan, bukti respons, dan batas interpretasi. Label C4 atau C5 dapat dicantumkan, tetapi disertai penjelasan tindakan. Pendekatan ini mengikuti penggunaan taksonomi sebagai alat penyelarasan tujuan dan tugas (Iowa State University CELT, n.d.). Penyusun juga perlu menentukan apakah peserta mengingat

informasi, menemukan hubungan, atau mengevaluasi kriteria. Satu soal dapat melibatkan beberapa tindakan; kategori utama dipilih berdasarkan tuntutan yang menentukan keberhasilan.

Distribusi kategori tidak menggantikan keseimbangan konstruk. Tes dengan banyak soal berlabel tinggi dapat mengabaikan kemampuan dasar yang diperlukan untuk membaca. Sebaliknya, pemeriksaan dasar tetap berguna untuk diagnosis. Tujuan asesmen menentukan campuran tugas, bukan tuntutan agar semua butir mempunyai kategori paling tinggi. Pengajar perlu menjelaskan mengapa suatu tugas ditempatkan dalam kisi-kisi dan informasi apa yang diberikan oleh hasilnya.

Tabel 4 Contoh kisi kisi operasional

Target	Stimulus	Tindakan dan bukti	Batas simpulan
Inferensi waktu	Pengumuman layanan	Memilih kunjungan dan menunjukkan petunjuk	Konsistensi temporal pada teks
Evaluasi informasi	Kebutuhan pengembalian buku	Menjelaskan kekurangan pengumuman	Kecukupan informasi bukan aturan sekolah
Revisi pesan	Pengumuman dan audiens	Menyusun tambahan dengan alasan	Kualitas revisi sesuai spesifikasi

Sumber: hasil analisis dan rancangan dalam artikel ini.

4.2 Rubrik dan umpan balik

Rubrik awal memisahkan ketepatan keputusan, relevansi bukti, dan hubungan inferensial. Setiap dimensi dapat diberi rentang nol sampai tiga. Skor tinggi menuntut kesesuaian keputusan dengan petunjuk serta alasan yang menjelaskan hubungan. Skor menengah menunjukkan sebagian unsur sudah tepat; skor rendah menunjukkan jawaban tanpa dasar atau tidak konsisten. Skala belum diuji dan tidak menjadi ambang kelulusan. Penilai memerlukan contoh respons untuk menyepakati makna deskriptor.

Umpan balik diarahkan pada dimensi yang lemah. Peserta yang memilih tepat tetapi tidak memberi bukti diminta menunjukkan bagian teks. Peserta yang memakai petunjuk benar namun menarik simpulan terlalu luas diminta membatasi klaim. Revisi disimpan sebagai bukti belajar formatif, sementara nilai respons awal dapat tetap digunakan jika tujuan pengukuran memerlukan kemandirian. Dengan demikian, peningkatan setelah bantuan tidak dicampur dengan kemampuan awal.

4.3 Fungsi digital dan keadilan akses

Sistem digital cukup menyediakan stimulus, ruang jawaban, kolom alasan, dan riwayat revisi. Fitur analitik hanya dipakai jika menjawab pertanyaan asesmen. Durasi, klik, dan pembukaan petunjuk memberi konteks, tetapi tidak diperlakukan sebagai ukuran langsung kualitas berpikir. Pengajar juga memeriksa tampilan Arab dari kanan ke kiri, keterbacaan harakat, dan kesesuaian ukuran layar. Kesalahan tampilan dapat menghasilkan masalah yang tidak berasal dari kemampuan peserta.

Penggunaan otomatisasi untuk menilai uraian membutuhkan kehati-hatian karena padanan redaksi dapat beragam. Sistem dapat membantu menemukan respons yang perlu ditelaah, tetapi penilaian penalaran tidak cukup berdasarkan kecocokan kata kunci. Aturan koreksi

harus dapat dijelaskan. Jika alat memberikan umpan balik yang berbeda kepada peserta, kondisi bantuan dicatat. Asesmen digital yang baik mempertahankan hubungan antara bukti dan keputusan, bukan hanya mempercepat rekapitulasi nilai.

4.4 Agenda validasi

Validasi dimulai dengan telaah hubungan tujuan, tugas, serta rubrik. Tahap berikutnya memeriksa proses respons melalui penjelasan peserta tentang cara mengerjakan. Dua penilai kemudian mencoba rubrik pada respons aktual untuk melihat konsistensi. Analisis kesulitan dan diskriminasi butir dilakukan setelah data tersedia; nilai statistik tidak dibuat dalam tahap konsep. Pengujian juga mempertimbangkan konsekuensi penggunaan, khususnya apabila skor dipakai untuk keputusan lebih luas daripada umpan balik kelas.

Studi Masrifah et al. (2025) memberi konteks penelitian latihan HOTS, tetapi kerangka baru tetap memerlukan pengujian sendiri. Penelitian selanjutnya dapat membandingkan format jawaban saja dengan format jawaban-plus-alasan pada tugas sepadan. Pertanyaan utamanya apakah bukti tambahan memperjelas diagnosis dan bukan hanya memperpanjang pengerjaan. Kriteria keberhasilan harus ditetapkan sebelum analisis. Artikel ini tidak melaporkan efek format tersebut.

4.5 Penggunaan hasil untuk keputusan pembelajaran

Hasil asesmen sebaiknya disajikan sebagai profil singkat, bukan hanya skor total. Dua peserta dapat memperoleh jumlah nilai sama melalui pola berbeda: peserta pertama memahami keputusan tetapi kurang menunjukkan bukti, sedangkan peserta kedua mampu menemukan petunjuk namun menarik simpulan yang terlalu luas. Keduanya memerlukan tindak lanjut berbeda. Profil dapat berisi contoh respons yang mendukung diagnosis serta satu tugas latihan yang sesuai. Pengajar tidak perlu mengubah profil menjadi label tetap tentang kecerdasan atau karakter. Yang dijelaskan adalah kualitas tindakan pada tugas yang telah dikerjakan.

Pada penggunaan formatif, pengajar dapat meminta peserta membandingkan alasan sebelum dan sesudah komentar. Perubahan yang tepat menjadi bahan diskusi tentang cara membaca bukti. Namun, pencatatan nilai tetap harus membedakan respons awal dan respons setelah bantuan. Jika skor revisi dipakai untuk penilaian perkembangan, aturan itu dinyatakan sejak awal. Jika yang diperlukan adalah kemampuan mandiri, respons setelah model tidak diperlakukan sebagai bukti yang setara. Kejelasan aturan membuat peserta memahami hubungan antara kesempatan memperbaiki dan keputusan penilaian.

Penggunaan hasil juga perlu mempertimbangkan apa yang belum diukur. Tugas pengumuman tidak mencakup seluruh jenis penalaran dalam sastra, percakapan, atau tulisan akademik. Pengajar dapat memakai beberapa stimulus dengan tuntutan relevan sebelum memperluas simpulan. Perluasan bukan sekadar menambah jumlah butir, tetapi menyediakan variasi yang mendukung klaim. Peserta diberi kesempatan melihat alasan penilaian dan menunjukkan jika alternatifnya masih sesuai dengan teks. Pemeriksaan semacam itu membantu memperbaiki soal yang ambigu. Prosedur merupakan rekomendasi penggunaan hasil, bukan laporan bahwa sebuah sistem digital sudah melaksanakannya.

5 Keterbatasan

Kajian bersifat naratif terarah dan menggunakan contoh buatan. Tidak ada sampel siswa, analisis respons, atau data perangkat. Kisi-kisi dan rubrik belum divalidasi. Oleh karena itu,

simpulan dibatasi pada hubungan argumentatif dalam desain. Naskah tidak menetapkan standar kemampuan HOTS untuk suatu jenjang. Penelaahan kedua penulis serta konfirmasi kontribusi, pendanaan, dan kebijakan pengungkapan bantuan AI diperlukan sebelum pengajuan.

6 Kesimpulan

Asesmen HOTS bahasa Arab perlu dibangun melalui hubungan klaim, tugas, bukti, dan interpretasi. Teks panjang atau kosakata sulit tidak cukup membuktikan tuntutan kognitif. Stimulus pendek dapat menuntut evaluasi jika peserta harus menyatakan alasan serta informasi yang belum diketahui. Lingkungan digital berguna untuk menyimpan proses, tetapi tidak menjamin kualitas berpikir. Kerangka, contoh, kisi-kisi, dan rubrik yang diusulkan memerlukan pengujian sebelum dipakai untuk menyimpulkan kemampuan peserta.

Daftar Pustaka

- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Iowa State University Center for Excellence in Learning and Teaching. (n.d.). Bloom's taxonomy. <https://celt.iastate.edu/prepare-and-teach/design-your-course/blooms-taxonomy/>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Masrifah, K., Ahsanuddin, M., & Ainin, M. (2025). The effectiveness of HOTS-based practice questions in enhancing students' critical thinking in Arabic learning. *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, dan Budaya Arab*, 8(2). <https://doi.org/10.17509/alsuniyat.v8i2.87847>
- Zubaidah, Ainin, M., Muassomah, Albantani, A. M., & Maimunah, I. (2022). Higher order thinking skills (HOTS) level instructional book of Arabic language at senior high school. *Arabiyat: Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban*, 9(2), 151–164. <https://doi.org/10.15408/a.v9i2.28184>