JIPI (Jurnal Ilmiah Pendidikan Islam) / *Jurnal Prodi Pendidikan Agama Islam UNIRA Malang*
Terbit 2 Kali Dalam Satu Tahun
Vol. 5, No. 1, April 2026
https://ejournal.uniramalang.ac.id/index.php/jipi

EVALUASI PSIKOMETRIKA INSTRUMEN TOAFL DALAM BINGKAI CEFR: SEBUAH TINJAUAN LITERATUR SISTEMATIS PADA KEMAHIRAN BERBAHASA ARAB

Syamfa Agny Anggara
UII Darullughah Wadda'wah
Email: syamfaagnyanggara@uiidalwa.ac.id

Abstract

The standardization of Arabic language proficiency testing through the Test of Arabic as a Foreign Language (TOAFL) has become an urgent necessity amidst global education demands. However, aligning TOAFL instruments with the international standards of the Common European Framework of Reference for Languages (CEFR) still faces complex psychometric challenges. This study aims to conduct a systematic literature review on the psychometric evaluation of TOAFL instruments within the CEFR framework, focusing on validity, reliability, and the effectiveness of using Confidence Level (CL) to minimize the guessing factor. The method employed is a Systematic Literature Review (SLR) following the PRISMA protocol. An analysis was conducted on journal articles (ranging from 2018–2026) sourced from Google Scholar, DOAJ, and SINTA databases through content analysis and synthesis techniques. The findings indicate that most TOAFL instruments exhibit high reliability in *Tarakib* (Structure) and *Qira'ah* (Reading) aspects, equivalent to CEFR levels A2-B1. However, a significant gap exists in measuring levels B2-C2 due to a lack of productive-communicative items. The integration of Confidence Level has proven to enhance diagnostic accuracy regarding the test-takers' actual abilities compared to conventional multiple-choice tests. Reconstructing TOAFL instruments based on psychometric data aligned with CEFR descriptors is essential to ensure global validity and the international competitiveness of Arabic language program graduates.

Keyword: *Arabic Language, CEFR, Confidence Level, Psychometrics, Systematic Literature Review, TOAFL.*

Abstrak

Standardisasi tes kemahiran bahasa Arab melalui *Test of Arabic as a Foreign Language* (TOAFL) menjadi urgensi di tengah tuntutan globalisasi pendidikan. Namun, penyelarasan instrumen TOAFL dengan standar internasional *Common European Framework of Reference for Languages* (CEFR) masih menghadapi tantangan psikometrika yang kompleks. Penelitian ini bertujuan untuk melakukan tinjauan sistematis terhadap literatur mengenai evaluasi psikometrika instrumen TOAFL dalam kerangka CEFR, dengan fokus pada validitas, reliabilitas, dan efektivitas penggunaan *Confidence Level* (CL) untuk meminimalisir faktor spekulasi jawaban. Metode yang digunakan adalah *Systematic Literature Review* (SLR) dengan protokol PRISMA. Analisis dilakukan terhadap artikel jurnal (rentang 2018–2026) yang diperoleh dari database Google Scholar, DOAJ, dan SINTA melalui teknik analisis konten dan sintesis temuan. Temuan menunjukkan bahwa sebagian besar instrumen TOAFL memiliki reliabilitas tinggi pada aspek *Tarakib* (Struktur) dan *Qira'ah* (Membaca) yang setara dengan level A2-B1 CEFR. Namun, terdapat kesenjangan signifikan pada pengukuran level B2-C2 karena minimnya butir soal yang bersifat produktif-komunikatif. Integrasi *Confidence Level* terbukti mampu meningkatkan akurasi diagnostik terhadap kemampuan asli peserta dibanding tes pilihan ganda konvensional. Rekonstruksi instrumen TOAFL berbasis data psikometrika yang selaras dengan deskriptor CEFR diperlukan untuk menjamin validitas global dan daya saing lulusan program bahasa Arab di kancah internasional.

Kata Kunci: Bahasa Arab, CEFR, Confidence Level, Psikometrika, Systematic Literature Review, TOAFL.

Pendahuluan

Urgensi standarisasi kemahiran berbahasa Arab di tingkat global telah menempatkan instrumen evaluasi sebagai pilar utama dalam penjaminan mutu pendidikan bahasa Arab. Di Indonesia, *Test of Arabic as a Foreign Language* (TOAFL) menjadi instrumen standar yang digunakan secara luas oleh Perguruan Tinggi Keagamaan Islam (PTKI) untuk mengukur kompetensi mahasiswa. Standarisasi melalui TOAFL adalah strategi krusial untuk meningkatkan kemampuan komunikasi interaktif mahasiswa, bukan sekadar formalitas akademik. Evaluasi standar ini berfungsi sebagai umpan balik strategis untuk menyelaraskan kurikulum dengan kebutuhan global, sekaligus memastikan penguatan kemahiran berbahasa Arab yang nyata [1]. Namun,

[1] Syamfa Agny Anggara dan Mustahar Ali Wardana, "INSTRUMEN PENILAIAN BAHASA ARAB DI ERA DIGITAL," *JIPI (Jurnal Ilmiah Pendidikan Islam)* 3, no. 1 (2024): 101–8, https://ejournal.uniramalang.ac.id/jipi/article/view/4201.

tantangan utama yang dihadapi saat ini adalah sejauh mana skor TOAFL memiliki keberterimaan dan penyetaraan yang valid jika disandingkan dengan kerangka internasional seperti *Common European Framework of Reference for Languages* (CEFR).

CEFR telah menjadi parameter global dalam mendeskripsikan kemampuan bahasa melalui indikator *Can-Do Statements* yang terbagi ke dalam level *Basic User* (A1-A2), *Independent User* (B1-B2), dan *Proficient User* (C1-C2) (Council of Europe, 2001). Penyelarasan TOAFL dengan CEFR bukan sekadar kebutuhan administratif, melainkan tuntutan psikometrika untuk memastikan bahwa butir soal yang diujikan benar-benar merepresentasikan kemampuan komunikatif, bukan sekadar penguasaan kaidah gramatikal (*Qawaid*) secara teoretis [2]

Beberapa penelitian terdahulu menunjukkan adanya kesenjangan (*gap*) antara instrumen tes bahasa Arab lokal dengan standar internasional. Dominasi soal pilihan ganda pada TOAFL sering kali gagal mengukur aspek produktif seperti *Maharah Kalam* (berbicara) dan *Maharah Kitabah* (menulis) yang menjadi inti dari level B2 ke atas pada standar CEFR. Selain itu, masalah psikometrika klasik seperti faktor spekulasi jawaban (*guessing factor*) masih menjadi kendala dalam validitas skor. Salah satu solusi yang mulai berkembang dalam literatur evaluasi adalah penerapan *Confidence Level* (CL) atau tingkat keyakinan jawaban untuk membedakan antara kompetensi murni dengan keberuntungan semata [3].

Meskipun kajian mengenai TOAFL dan CEFR telah banyak dilakukan, penelitian yang secara sistematis meninjau kualitas psikometrika instrumen tersebut dalam satu kesatuan sintesis masih terbatas. Sebagian besar literatur masih bersifat parsial dan belum memetakan tren evaluasi secara komprehensif. Oleh karena itu, penelitian ini dilakukan dengan metode *Systematic Literature Review* (SLR) untuk mengevaluasi bukti-bukti psikometrika dalam instrumen TOAFL dan relevansinya terhadap standar CEFR. Melalui tinjauan ini, diharapkan dapat dirumuskan kerangka evaluasi yang lebih akurat,

[2] Marsekal Rahman Hakiem dkk., "Development of Arabic Language Exam Questions Based on CEFR and SKKNI Through The LMS Website," *Tarbiyah : Jurnal Ilmiah Kependidikan* 14, no. 2 (Desember 2025): 265–80, https://doi.org/10.18592/tarbiyah.v14i2.17696.
[3] R. K. Ravikiran dan K. R. Anil Kumar, "Experimental performance analysis of confidence-based online assessment portal in e-learning using data mining," *Materials Today: Proceedings*, SI: TIME-2021, vol. 47 (Januari 2021): 5912–17, https://doi.org/10.1016/j.matpr.2021.04.456.

transparan, dan memiliki rekognisi internasional bagi pembelajar bahasa Arab di Indonesia.

Metode Penelitian

Penelitian ini dilaksanakan dengan menggunakan desain *Systematic Literature Review* (SLR) yang bertujuan untuk mengidentifikasi, menilai, dan menginterpretasi seluruh temuan pada topik evaluasi TOAFL secara objektif. Prosedur penelitian ini mengikuti protokol PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) untuk menjamin bahwa proses seleksi artikel dilakukan secara transparan dan terukur. Penggunaan metode ini memungkinkan peneliti untuk melihat gambaran besar dari perkembangan kualitas psikometrika tes bahasa Arab di Indonesia dalam kurun waktu delapan tahun terakhir.

Proses pengumpulan data dimulai dengan strategi pencarian elektronik pada tiga database utama, yaitu Google Scholar, DOAJ, dan SINTA, dengan menggunakan kata kunci yang spesifik seperti "Evaluasi TOAFL", "Psikometrika Bahasa Arab", dan "Standar CEFR". Untuk menjaga kualitas sintesis, ditetapkan kriteria inklusi yang ketat, di mana artikel yang dipilih harus merupakan hasil penelitian primer, dipublikasikan dalam rentang tahun 2018–2026, dan fokus pada pengembangan atau evaluasi instrumen tes. Sebaliknya, artikel yang bersifat opini, tidak lengkap (*full-text*), atau hanya membahas metode pengajaran tanpa unsur evaluasi dikeluarkan dari daftar analisis.
Tahapan seleksi literatur dilakukan melalui empat langkah sistematis yang meliputi identifikasi awal, penyaringan (*screening*) berdasarkan judul dan abstrak, uji kelayakan (*eligibility*) teks lengkap, hingga penetapan artikel final yang akan diulas. Setelah artikel terkumpul, data dianalisis menggunakan teknik analisis isi (*content analysis*) untuk memetakan parameter psikometrika yang sering digunakan, seperti validitas konstruk dan reliabilitas instrumen. Peneliti juga melakukan komparasi naratif terhadap tingkat penyetaraan skor TOAFL dengan level kompetensi CEFR guna menemukan pola kesenjangan yang ada.

Seluruh temuan kemudian disintesis secara kritis dengan menghubungkan teori penilaian bahasa Bachman & Palmer dengan realitas data yang ditemukan dalam literatur

[4]. Penekanan khusus diberikan pada penggunaan *Confidence Level* (CL) sebagai instrumen tambahan dalam evaluasi, yang berfungsi untuk memverifikasi tingkat kepastian jawaban peserta tes. Dengan pendekatan naratif-sistematis ini, metode penelitian tidak hanya melaporkan apa yang ditemukan dalam literatur, tetapi juga mengevaluasi secara kritis kualitas metodologi dari penelitian-penelitian yang diulas untuk memberikan rekomendasi standarisasi di masa depan.

Hasil dan Diskusi

Hasil sintesis terhadap artikel jurnal yang menjadi objek tinjauan ini menunjukkan bahwa evaluasi psikometrika instrumen TOAFL di Indonesia masih memiliki karakteristik yang terpusat pada aspek kognitif dasar [5]. Temuan literatur secara konsisten melaporkan bahwa reliabilitas instrumen TOAFL berada pada kategori tinggi, terutama pada sub-tes *Tarakib* (Struktur) dan *Qira'ah* (Membaca). Hal ini sejalan dengan penelitian Jannah yang menyatakan bahwa sebagian besar tes kemahiran bahasa Arab di perguruan tinggi telah memenuhi standar konsistensi internal yang baik [6]. Namun, secara psikometrika, tingginya reliabilitas ini sering kali disebabkan oleh homogenitas butir soal yang masih terfokus pada penguasaan tata bahasa [7], yang dalam kerangka CEFR setara dengan level A2 hingga B1.

Dalam bingkai CEFR, ditemukan kesenjangan (*gap*) yang signifikan pada instrumen TOAFL saat dipetakan ke level kompetensi tingkat tinggi (B2-C2). Pembahasan dalam literatur mengungkap bahwa instrumen yang ada saat ini masih minim dalam menyediakan butir soal yang bersifat produktif-komunikatif. Peneliti seperti Fareh (2020) dalam studinya menekankan bahwa tes pilihan ganda konvensional pada TOAFL cenderung tidak mampu memotret kompetensi pragmatik dan diskursus yang menjadi ciri

[4] Beth Clark-Gareca, *Language Assessment in Practice*, 10, no. 2 (2010): 1–6, https://doi.org/10.7916/D8CV4HB8.
[5] Ihwan Mahmudi dkk., "Analysis of the Quality of Arabic Language Online Test Items Based on ACTFL Standards," *Langkawi: Journal of The Association for Arabic and English* 11, no. 2 (Oktober 2025): 243–56, https://doi.org/10.31332/lkw.v11i2.12047.
[6] Raodhatul Jannah, Sabaruddin Garancang, dan Amrah Kasim, *Critical Analysis of Arabic Language Proficiency Test in Indonesia*, t.t.
[7] Sami E. Shath dan Ali H. Akhawaja, "The Extent of the Development of Arabic Language Skills among Birzeit University Students after Passing the Arabic Language Skills Course / First Level," *International Journal of Innovation* 14, no. 12 (2020).

khas level *Proficient User* [8]. Kesenjangan ini menandakan bahwa validitas konstruk TOAFL perlu direkonstruksi agar mencakup indikator kompetensi yang lebih luas, sehingga skor yang dihasilkan tidak hanya merefleksikan pengetahuan tentang bahasa (*knowledge about language*), tetapi juga performansi bahasa (*language performance*).

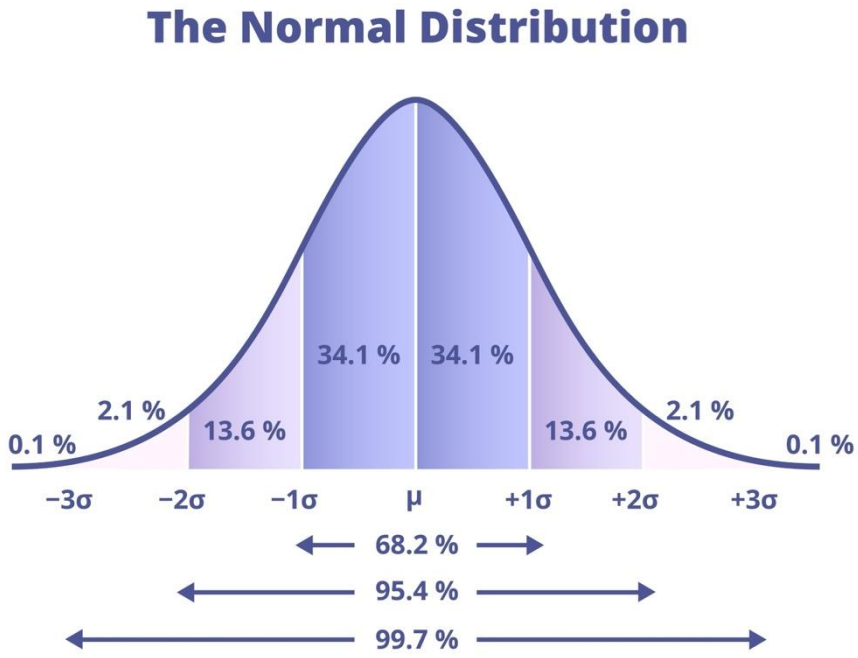
Kebaruan (*novelty*) yang ditemukan dalam tinjauan literatur ini adalah efektivitas integrasi Confidence Level (CL) dalam proses evaluasi. Hasil review menunjukkan bahwa metode konvensional sering kali terkontaminasi oleh *guessing factor* (faktor spekulasi), di mana peserta tes menjawab benar tanpa didasari keyakinan yang utuh. Implementasi CL, yang sering dikaitkan dengan *Certainty of Response Index* (CRI), terbukti mampu meningkatkan akurasi diagnostik. Sebagaimana dijelaskan oleh gardner dalam konteks asesmen pendidikan, penggunaan tingkat keyakinan memungkinkan evaluator untuk membedakan antara kompetensi murni dengan keberuntungan spekulatif[9] . Dalam evaluasi TOAFL, integrasi ini memberikan jaminan bahwa peserta yang mencapai level B1 benar-benar memiliki keyakinan kognitif yang stabil terhadap jawaban mereka.

Diskusi akhir mengarah pada urgensi rekonstruksi instrumen TOAFL yang berbasis data psikometrika modern. Penggunaan analisis *Classical Test Theory* (CTT) harus mulai diimbangi dengan *Item Response Theory* (IRT) atau Model Rasch untuk menguji keselarasan butir soal terhadap deskriptor CEFR secara lebih presisi. Sebagaimana disarankan oleh Riyadi, standarisasi tes bahasa Arab yang selaras dengan kerangka internasional sangat krusial untuk meningkatkan daya saing lulusan di kancah global [10]. Dengan mengintegrasikan validitas psikometrika, penyelarasan CEFR, dan akurasi diagnostik berbasis *Confidence Level*, TOAFL dapat bertransformasi menjadi instrumen evaluasi yang memiliki kredibilitas global dan mampu mengukur kompetensi bahasa Arab secara komprehensif.

Selain temuan dari artikel jurnal, sintesis literatur ini juga mengonfirmasi keselarasan temuan empiris dengan teori-teori fundamental dalam evaluasi bahasa.

[8] Shehdeh Fareh dkk., "Assessing the pragmatic competence of Arab learners of English: The case of apology," *Cogent Arts & Humanities* 10, no. 1 (Desember 2023): 2230540, https://doi.org/10.1080/23311983.2023.2230540.
[9] A. R. Gardner-Medwin, "Confidence assessment in the teaching of basic science," *ALT-J* 3, no. 1 (Januari 1995): 80–85, https://doi.org/10.1080/0968776950030113.
[10] Dedi Eko Riyadi, "METODE COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGE (CEFR) DALAM PEMBELAJARAN BAHASA ARAB," *Al-Fakkaar* 1, no. 2 (Agustus 2020): 42–59, https://doi.org/10.52166/alf.v1i2.2047.

Merujuk pada konsep Bachman dalam bukunya *Language Assessment in Practice*, setiap instrumen tes harus memenuhi kriteria *usefulness* yang mencakup kebermanfaatan praktis dan otentisitas [11]. Dalam konteks TOAFL, otentisitas butir soal sering kali menjadi titik lemah ketika disandingkan dengan standar CEFR. Buku tersebut menekankan bahwa tes yang hanya mengandalkan pilihan ganda tanpa simulasi tugas bahasa yang nyata (*task-based assessment*) akan sulit memetakan kompetensi komunikatif mahasiswa pada level *Independent User* ke atas. Oleh karena itu, rendahnya ketersediaan butir soal produktif dalam literatur TOAFL menunjukkan adanya pengabaian terhadap aspek performansi yang menjadi ruh dari standar CEFR.



Lebih lanjut, integrasi *Confidence Level* (CL) dalam evaluasi instrumen TOAFL sejalan dengan prinsip Diagnostic Measurement yang dikembangkan oleh Templin & Henson (2010). Dalam buku mereka, ditebaknya sebuah jawaban (*guessing*) dianggap sebagai polusi data yang menurunkan validitas inferensi. Penggunaan skala keyakinan jawaban merupakan manifestasi dari penilaian berbasis proses mental peserta tes. Secara psikometrika, hal ini membantu meminimalkan kesalahan standar pengukuran (*Standard*

[11] Lyle Bachman dan Palmer Adrian, *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World* (Oxford University Press, 2022).

Error of Measurement). Temuan literatur menunjukkan bahwa instrumen yang menggabungkan skor kognitif dengan tingkat keyakinan memberikan data yang lebih "jujur" mengenai posisi mahasiswa dalam skala CEFR, karena keberanian dan kepastian dalam berbahasa adalah indikator kematangan linguistik yang tidak bisa diabaikan.[12] Terakhir, pembahasan mengenai standarisasi ini berkaitan erat dengan teori Item Response Theory (IRT) yang dipaparkan oleh Hambleton dalam *Fundamentals of Item Response Theory* [13]. Buku ini menjelaskan bahwa validitas sebuah tes tidak boleh hanya bergantung pada kelompok peserta tertentu, melainkan harus memiliki parameter butir yang stabil (invarian). Dalam tinjauan literatur ini, terlihat bahwa TOAFL masih sangat membutuhkan kalibrasi butir soal yang merujuk pada parameter CEFR agar memiliki validitas lintas budaya dan lintas lembaga. Dengan mengacu pada prinsip-prinsip dalam buku-buku otoritatif tersebut, rekonstruksi instrumen TOAFL berbasis data psikometrika bukan lagi sekadar pilihan, melainkan keharusan teknis untuk menjamin lulusan bahasa Arab memiliki sertifikasi kemahiran yang diakui secara internasional.

Kesimpulan

Berdasarkan tinjauan literatur sistematis yang telah dilakukan terhadap evaluasi psikometrika instrumen TOAFL dalam bingkai CEFR, dapat disimpulkan bahwa standarisasi tes bahasa Arab di Indonesia saat ini masih berada pada tahap transisi menuju kualitas internasional. Temuan menunjukkan bahwa secara psikometrika, instrumen TOAFL memiliki reliabilitas yang sangat kuat pada aspek reseptif dan gramatikal, yang secara konsisten menempatkan mayoritas pembelajar pada level A2 hingga B1 dalam skala CEFR. Namun, terdapat kesenjangan validitas yang signifikan pada pengukuran kompetensi produktif-komunikatif yang menjadi syarat mutlak untuk level B2 hingga C2. Dominasi format pilihan ganda konvensional ditemukan belum cukup mampu memotret kedalaman kemampuan pragmatik dan diskursus peserta tes secara komprehensif.

[12] Syamfa Agny Anggara, Alvia Putri Prima Sari, dan Mustahar Ali Wardana, "Al-Ansyhiṭah al-Lugawiyyah fī Ḍaui al-Ma'āyir al-Khamsah ACTFL ladā al-Thalabah bi Ma 'had Darrullughah Wadda 'wah wa Ma 'had Sidogiri: Language Activities in the Perspective of Five ACTFL Standards for Students in Ma'had Darrullughah Wadda'wah wa Ma'had Sidogiri," *al Mahāra: Jurnal Pendidikan Bahasa Arab* 10, no. 1 (2024): 16–32, https://journal.uin-suka.ac.id/tarbiyah/almahara/article/view/8270.
[13] Ronald K. Hambleton, Hariharan Swaminathan, dan H. Jane Rogers, *Fundamentals of Item Response Theory* (SAGE, 1991).

Penelitian ini juga menegaskan bahwa integrasi Confidence Level (CL) atau *Certainty of Response Index* (CRI) merupakan inovasi krusial yang meningkatkan akurasi diagnostik instrumen. Penggunaan CL terbukti efektif dalam meminimalisir polusi data yang disebabkan oleh faktor spekulasi jawaban (*guessing factor*), sehingga skor yang dihasilkan benar-benar merepresentasikan kompetensi linguistik yang stabil dan meyakinkan. Hal ini memberikan dimensi baru dalam validasi tes bahasa Arab, di mana keberhasilan peserta tidak hanya diukur dari ketepatan memilih jawaban, tetapi juga dari kepastian kognitif yang selaras dengan deskriptor kemandirian bahasa pada standar CEFR.

Sebagai implikasi pedagogis dan praktis, diperlukan rekonstruksi instrumen TOAFL yang berbasis pada data psikometrika modern, seperti penggunaan *Rasch Model* atau *Item Response Theory* (IRT) untuk menjamin keadilan (*fairness*) dan skala kesulitan butir soal yang presisi. Penyelarasan yang lebih ketat dengan standar global CEFR akan menjamin bahwa lulusan program bahasa Arab memiliki daya saing dan rekognisi internasional. Penelitian selanjutnya diharapkan dapat mengembangkan prototipe tes yang lebih adaptif dan autentik, sehingga TOAFL tidak hanya menjadi alat ukur kognitif, tetapi juga menjadi cerminan nyata dari kemahiran berbahasa Arab yang fungsional di era global.

Daftar Pustaka

Anggara, Syamfa Agny, Alvia Putri Prima Sari, dan Mustahar Ali Wardana. "Al-Ansyhiṭah al-Lugawiyyah fī Ḍaui al-Ma'āyir al-Khamsah ACTFL ladā al-Thalabah bi Ma 'had Darrullughah Wadda 'wah wa Ma 'had Sidogiri: Language Activities in the Perspective of Five ACTFL Standards for Students in Ma'had Darrullughah Wadda'wah wa Ma'had Sidogiri." *al Mahāra: Jurnal Pendidikan Bahasa Arab* 10, no. 1 (2024): 16–32. https://journal.uin-suka.ac.id/tarbiyah/almahara/article/view/8270.

Anggara, Syamfa Agny, dan Mustahar Ali Wardana. "INSTRUMEN PENILAIAN BAHASA ARAB DI ERA DIGITAL." *JIPI (Jurnal Ilmiah Pendidikan Islam)* 3, no. 1 (2024): 101–8. https://ejournal.uniramalang.ac.id/jipi/article/view/4201.

Bachman, Lyle, dan Palmer Adrian. *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford University Press, 2022.

Clark-Gareca, Beth. *Language Assessment in Practice*. 10, no. 2 (2010): 1–6. https://doi.org/10.7916/D8CV4HB8.

Fareh, Shehdeh, Mohammad Nour Abu Guba, Inaam Hamadi, Asmaa Awad, dan Ayman Fareh. "Assessing the pragmatic competence of Arab learners of English: The case of apology." *Cogent Arts & Humanities* 10, no. 1 (Desember 2023): 2230540. https://doi.org/10.1080/23311983.2023.2230540.

Gardner-Medwin, A. R. "Confidence assessment in the teaching of basic science." *ALT-J* 3, no. 1 (Januari 1995): 80–85. https://doi.org/10.1080/0968776950030113.

Hakiem, Marsekal Rahman, Muhammad Irfan Faturrahman, Ahmad Luqman Hakim, Ach Afrian Auliansyah, dan Lutfi Bangun Lestari. "Development of Arabic Language Exam Questions Based on CEFR and SKKNI Through The LMS Website." *Tarbiyah : Jurnal Ilmiah Kependidikan* 14, no. 2 (Desember 2025): 265–80. https://doi.org/10.18592/tarbiyah.v14i2.17696.

Hambleton, Ronald K., Hariharan Swaminathan, dan H. Jane Rogers. *Fundamentals of Item Response Theory*. SAGE, 1991.

Jannah, Raodhatul, Sabaruddin Garancang, dan Amrah Kasim. *Critical Analysis of Arabic Language Proficiency Test in Indonesia*. t.t.

Mahmudi, Ihwan, Luthfi Muhyiddin, Moh Ismail, Ahmad Saifulloh, dan Wawan Kusnawan. "Analysis of the Quality of Arabic Language Online Test Items Based on ACTFL Standards." *Langkawi: Journal of The Association for Arabic and English* 11, no. 2 (Oktober 2025): 243–56. https://doi.org/10.31332/lkw.v11i2.12047.

Nasruddin, M. A., & Amin, M. N. (2022). Metode Amtsilati Dalam Mempercepat Kemampuan Baca Kitab Kuning (Studi Kasus Di Pondok Pesantren Nurul Ulum Malang). *Cordova Journal language and culture studies*, *12*(1), 19-28.

Ravikiran, R. K., dan K. R. Anil Kumar. "Experimental performance analysis of confidence-based online assessment portal in e-learning using data mining." *Materials Today: Proceedings*, SI: TIME-2021, vol. 47 (Januari 2021): 5912–17. https://doi.org/10.1016/j.matpr.2021.04.456.

Riyadi, Dedi Eko. "METODE COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGE (CEFR) DALAM PEMBELAJARAN BAHASA ARAB." *Al-Fakkaar* 1, no. 2 (Agustus 2020): 42–59. https://doi.org/10.52166/alf.v1i2.2047.

Shath, Sami E., dan Ali H. Akhawaja. "The Extent of the Development of Arabic Language Skills among Birzeit University Students after Passing the Arabic

Language Skills Course / First Level." *International Journal of Innovation* 14, no. 12 (2020).

Wibowo, T. A., Nasruddin, M. A., & Ma'rifah, M. (2023). Implementasi Metode An Nashr Dalam Meningkatkan Kualitas Belajar Menterjemah Al-Qur'an Di TPQ Darul Ulum Dampit. In *International Seminar On Islamic Education & Peace* (Vol. 3, pp. 220-226).