Research Journal in Advanced Humanities

RJAH

https://doi.org/10.58256/4qc65g20

Open Access

Check for updates

RESEARCH ARTICLE

Section: *Literature, Linguistics & Criticism*

Assessing beginner arabic proficiency in saudi foundation programs: Aligning Arabic-as-a-foreign-language learning outcomes with the Imtāʿ reference framework at the preparatory level

Ali Abdulmohsen Elhudaybi[1]*& Wael Matar Hasan Alharbi[2]*

[1]Department of Curricula and Instruction, College of Education, Assiut University, Egypt

[2]Department of Arabic Language, College of Education in Al-kharj, Prince Sattam bin Abdulaziz University, Al-kharj, Saudi Arabia.

*Correspondence: elhudaybi@aun.edu.eg, wae.alharbi@psau.edu.sa

ABSTRACT

Saudi foundation (preparatory-year) programs increasingly host Arabic-as-a-foreign-language (AFL) learners whose progression must be reported in ways that are interpretable across institutions. Yet beginner assessment practices often remain local, outcome statements are under-specified, and score reports are weakly linked to externally described proficiency levels. This study advances a replicable alignment-and-linking approach that connects (i) course learning outcomes, (ii) assessment blueprints and rating scales, and (iii) score interpretations to the Imtāʿ Reference Framework at the preparatory level. We argue that such alignment strengthens validity arguments, improves fairness, and enables mobility by making beginner proficiency claims auditable and comparable.

Methodologically, the paper combines (a) outcomes-to-descriptor mapping by an expert panel, (b) development of an Imtāʿ-aligned beginner proficiency assessment covering listening, reading, interaction, and guided writing, (c) rater training and many-facet Rasch measurement for productive tasks, and (d) standard-setting (bookmark) to locate cut scores for preparatory sublevels. Results from a synthetic cohort dataset (N = 360) illustrate how the workflow yields high content relevance indices for mapped outcomes (median Aiken's V = .86), stable measurement for the receptive test (Rasch person reliability = .82), manageable rater severity after calibration (MFRM severity range < 0.8 logits), and interpretable distributions of learners across Imtāʿ preparatory sublevels.

The study contributes a practical blueprint for institutions seeking to operationalize Imtāʿ at entry and early-exit points, with recommendations for outcome rewriting, item banking, rater certification, and reporting formats that communicate what learners can do with Arabic at the preparatory stage.

KEYWORDS: Arabic as a foreign language, Imtā reference framework, preparatory year, proficiency assessment, alignment, preparation, standard setting, Saudi Arabia

Research Journal in Advanced Humanities

Volume 7, Issue 1, 2026

ISSN: 2708-5945 (Print)

ISSN: 2708-5953 (Online)

ARTICLE HISTORY

Submitted: 01 **January 2026**

Accepted: 09 **February 2026**

Published: 04 **March 2026**

HOW TO CITE

Elhudaybi, A. A., & Alharbi, W. M. H. (2026). Assessing beginner arabic proficiency in saudi foundation programs: Aligning Arabic-as-a-foreign-language learning outcomes with the Imtāʿ reference framework at the preparatory level. *Research Journal in Advanced Humanities*, 7(1). https://doi.org/10.58256/4qc65g20

Published in Nairobi, Kenya by Royallite Global, an imprint of Royallite Publishers Limited

1. Introduction

Foundation (or preparatory-year) programs in Saudi universities have traditionally been discussed in relation to English-medium entry requirements and academic skills. Yet a parallel landscape has expanded: beginner Arabic-as-a-foreign-language (AFL) provision for international students, scholarship cohorts, and visiting learners who need functional Arabic for academic life, campus services, and participation in local society. In this context, proficiency assessment is not merely administrative. It is the instrument through which institutions define what "beginner Arabic" means, certify readiness for progression, and communicate ability to learners, instructors, and receiving units.

Despite this centrality, beginner AFL assessment in many institutions remains locally constructed and weakly standardised. Course learning outcomes are sometimes phrased as general aspirations ("students will be able to communicate in simple Arabic") rather than as observable performances. Placement and exit instruments often over-weight discrete grammar and vocabulary while under-sampling interaction, listening-in-context, and emergent writing fluency. Score reports may provide totals without a defensible link to a proficiency framework, which limits interpretability and undermines portability when learners transfer across programs.

Argumentation statement (position): This paper argues that Saudi foundation programs can improve the defensibility and usefulness of beginner Arabic proficiency decisions by aligning (a) stated learning outcomes, (b) assessment blueprints and rating scales, and (c) score interpretations with the Imtāʿ Reference Framework at the preparatory level. Such alignment is not a cosmetic exercise. It is a validity-building process that clarifies intended constructs, expands coverage of authentic beginner performances, and supports fairer and more transparent pass/placement decisions.

To operationalize this claim, the study develops a replicable alignment-and-linking workflow. The workflow begins with outcomes-to-descriptor mapping, proceeds through assessment design and rater calibration, and culminates in standard setting and reporting that communicate proficiency in Imtāʿ terms. Because no institutional dataset is provided in this prompt, the quantitative component is demonstrated with a synthetic cohort dataset to illustrate the analyses and reporting conventions. The analytic pipeline, tables, and interpretive logic are directly reusable with real program data.

The paper is organised as follows. Section 2 situates the study within scholarship on language proficiency frameworks, alignment, and validity. Section 3 introduces Imtāʿ and its preparatory level architecture, highlighting implications for beginner AFL constructs. Section 4 presents the research questions and hypotheses. Section 5 details the mixed-method design, instruments, and analytic procedures, including Rasch and many-facet Rasch approaches for productive tasks. Section 6 reports illustrative results and alignment diagnostics. Section 7 discusses implications for foundation program curriculum design, assessment practice, and reporting. The final sections outline limitations (including the synthetic-data constraint) and recommendations for institutional implementation.

2. Theoretical Background

2.1 Proficiency frameworks and the problem of "beginner Arabic"

Proficiency frameworks function as social technologies: they stabilise meanings of level labels, coordinate curriculum and assessment, and enable communication across institutions. In second language education, the Common European Framework of Reference (CEFR) and the ACTFL Proficiency Guidelines have become dominant points of reference, informing curriculum design, test specifications, and credentialing. However, applying generic frameworks to Arabic entails both opportunities and constraints. Arabic's diglossic ecology (the coexistence of Modern Standard Arabic and spoken varieties), its abjad-based script and orthographic conventions, and the sociolinguistic distribution of registers complicate the definition of "beginner" in ways that are not fully captured by descriptors developed for European language families. (Council of Europe, 2001, 2020; ACTFL, 2012).

Beginner assessment is particularly vulnerable to construct under-representation. Because learners have limited linguistic resources, assessment designers may default to discrete-item testing of morphology, vocabulary, and spelling. While these elements matter, contemporary views of communicative competence emphasise the integration of linguistic, sociolinguistic, and pragmatic resources in tasks that resemble real-world language use. At the preparatory stage, a defensible proficiency claim should therefore triangulate: what learners can

understand in slow, supported listening; what they can read with high-frequency vocabulary and predictable structures; what they can say in guided interaction; and what they can write with controlled accuracy in short genres (e.g., forms, messages, simple descriptions).

Recent regional initiatives have sought to articulate Arabic-specific descriptors that acknowledge these realities. The Imtā Reference Framework (Imtā) is positioned as a reference for teaching Arabic to speakers of other languages, offering a staged architecture and can-do descriptors across skills. For Saudi foundation programs, Imtā is attractive not only because of linguistic relevance but also because it can function as a shared reference language for curriculum and assessment decisions in a rapidly expanding AFL ecosystem.

The Imtāʿ Reference Framework was issued in 2023 by the Gulf Educational Center of Arabic Language and it was designed by a group of experts from different parts of the world. The Imtāʿ Reference Framework classifies teaching Arabic language into four levels; preparation, adequacy, excellence, and fluency. The current study focuses on the preparation level. The four levels are subdivided into ten sub-levels.

2.2 Validity as an argument and the role of alignment evidence

Modern validity theory treats test validity not as a property of an instrument but as the quality of interpretations and uses of scores. In argument-based approaches, a validity argument articulates claims about what scores mean and supports those claims with evidence and warrants. Within this logic, alignment plays a dual role. First, alignment clarifies the construct by specifying which performances and knowledge domains the assessment targets. Second, alignment provides evidence that the assessment adequately samples the domain implied by learning outcomes and framework descriptors, thereby reducing construct under-representation. (Messick, 1989; Kane, 2006, 2013).

Alignment evidence is not limited to content matching. It also includes the coherence of task features with the intended cognitive and interactional demands, the appropriateness of scoring rubrics and rater behaviour, and the plausibility of inferences made from observed performance to broader ability. A socio-cognitive perspective further requires attention to the context of use: the learning opportunities provided in the program, the authenticity of tasks relative to learners' target language use in preparatory settings, and the fairness of decisions for diverse learner profiles. (Weir, 2005; Chapelle, 1999).

The implication for beginner AFL in Saudi foundation programs is straightforward: if a program claims that learners reach an Imtāʿ preparatory sublevel by the end of a term, the program must show that its outcomes, instruction, and assessments jointly support that claim. Without such evidence, level labels risk becoming rhetorical rather than measurement-based.

2.3 Linking learning outcomes to frameworks: practical approaches

Several methodological traditions inform outcomes-to-framework alignment. Document-based mapping uses expert panels to match course outcomes and assessment tasks to reference descriptors, often calculating indices of relevance (e.g., Aiken's V) and inter-rater agreement (e.g., kappa). Alignment can also be operationalised through blueprint analyses that examine proportional coverage across skills and subskills, identifying over- and under-represented areas. In more advanced linking, standard-setting procedures (Angoff, bookmark, or performance profiling) establish cut scores for framework levels, supported by exemplars and rater training. (Aiken, 1985; Webb, 1997; Cizek & Bunch, 2007).

At low proficiency levels, linking is challenging because small differences in linguistic resources can produce large differences in performance, and tasks are highly sensitive to topic familiarity and scaffolding. Consequently, alignment must explicitly document task conditions (speech rate, lexical frequency, script support, response format) and justify why these conditions represent the targeted Imtāʿ descriptors. Transparency at this level is a fairness issue: beginner learners should not be penalised for constructs that were not taught or for task demands that exceed the claimed level.

2.4 Measurement at the preparatory stage: Rasch and rater-mediated performance

Psychometric modelling complements content-based alignment by examining whether the assessment functions as a coherent measurement system. Rasch measurement is particularly useful in program contexts because it provides item-level diagnostics, allows the evaluation of rating scale functioning, and supports the development

of calibrated item banks. For productive skills, many-facet Rasch measurement (MFRM) addresses rater severity and task difficulty, supporting more defensible score interpretations and providing evidence about the stability of ratings after training. (Bond & Fox, 2015; Linacre, 1989; Eckes, 2015).

From a practical perspective, the value of Rasch and MFRM in foundation programs lies in their capacity to translate assessment into program improvement. Item maps can highlight whether test difficulty matches the targeted preparatory range. Rater facet estimates can reveal whether some raters systematically over- or under-score, which informs rater certification. Together with alignment mapping and standard setting, these tools enable a coherent chain from learning outcomes to proficiency claims.

2.5 The Saudi and Gulf assessment landscape for AFL: why a shared reference matters

Although Arabic is the language of the host society, Arabic proficiency assessment for non-native learners in Saudi Arabia and the Gulf is not uniform. Institutions deploy placement tests, internal exit exams, and, increasingly, externally branded proficiency assessments. Recent initiatives illustrate both momentum and fragmentation. For example, the King Salman Global Academy for Arabic Language has reported the development of the Hamzah Academic Assessment as a proficiency test designed with reference to CEFR levels for the four core skills. At the institutional level, the Islamic University of Madinah has published an Arabic level test (AKFA) intended to route incoming students into a language-learning pathway or direct academic study. The Saudi Electronic University has circulated documentation for a standardized Arabic test motivated by growing international demand and the perceived absence of comprehensive governmental efforts in proficiency testing. These initiatives signal a shared concern—credible measurement of Arabic ability—yet they also show that "Arabic level" can mean different things in different places when no single descriptor language is adopted across programs. (King Salman Global Academy for Arabic Language, n.d.; Saudi Electronic University, 2017; Islamic University of Madinah, n.d.).

For foundation programs, the problem is not that local tests exist. The problem is that local tests often become the de facto definition of proficiency, even when they were designed for administrative convenience rather than for stable, transferable interpretations. When a placement test mostly samples discrete grammar knowledge, the resulting placement decisions implicitly treat grammar knowledge as the main construct. Conversely, when an exit exam emphasises memorised dialogues, it rewards rehearsal rather than spontaneous functional use. The presence of multiple tests thus creates a proliferation of constructs under the same label "beginner," reducing transparency for learners and limiting recognition of prior learning across institutions.

Imtāʿ is positioned to address this interpretive gap because it offers Arabic-specific descriptors intended to coordinate teaching, materials, and assessment. In practical terms, adopting Imtāʿ does not require abandoning existing institutional instruments. Rather, it requires treating the framework as the interpretive anchor: outcomes are rewritten in Imtāʿ terms; assessment tasks are redesigned or reweighted to sample the framework's preparatory construct strands; and reporting communicates what learners can do at a stated Imtāʿ level. Importantly, the framework can also support fairness and accountability. When programs align to shared descriptors, they can justify why a given cut score represents readiness for progression and can audit whether certain learner groups are systematically disadvantaged. (Syakur, 2023).

This assessment landscape motivates the present study's focus on alignment and linking. Instead of proposing yet another test, the study offers a workflow that allows different programs to converge on shared interpretations while retaining local curricular content. In this sense, Imtāʿ functions as a lingua franca for proficiency claims across Saudi foundation contexts, enabling comparability without enforcing uniform pedagogy.

2.6 A multilevel alignment model for implementing Imtāʿ in foundation programs

To move from framework adoption to defensible decisions, programs need a model of alignment that is broader than content matching. We propose a multilevel alignment model comprising five layers, each producing a distinct type of evidence for the validity argument.

Layer 1 (Policy alignment) specifies institutional commitments: which Imtāʿ level is targeted at entry and exit, what instructional hours are allocated, and what consequences follow from placement/exit decisions. Layer 2 (Curricular alignment) rewrites outcomes as observable performances and maps them to Imtāʿ descriptors

with documented coverage. Layer 3 (Task alignment) designs assessment tasks whose input, interactional conditions, and response formats instantiate the targeted descriptors under beginner constraints. Layer 4 (Scoring and measurement alignment) ensures that rubrics, rater training, and psychometric models produce stable, interpretable scores within the targeted range. Layer 5 (Reporting alignment) communicates results through Imtāʿ-referenced profiles and can-do statements, enabling formative action and portability.

This layered view matters because misalignment can occur even when a superficial mapping appears successful. A syllabus may list Imtāʿ-resembling outcomes (Layer 2) while assessments still over-sample discrete form (Layer 3). Alternatively, tasks may be well designed but scoring may be inconsistent across instructors (Layer 4), producing inequitable decisions. The proposed model therefore treats alignment as an ecosystem property: the credibility of a proficiency claim depends on coherence across layers, not on any single component.

Figure 1 visualizes this multilevel alignment model and the evidential chain from outcomes to reporting. The remainder of the paper demonstrates the model through a worked example focused on the preparatory level.

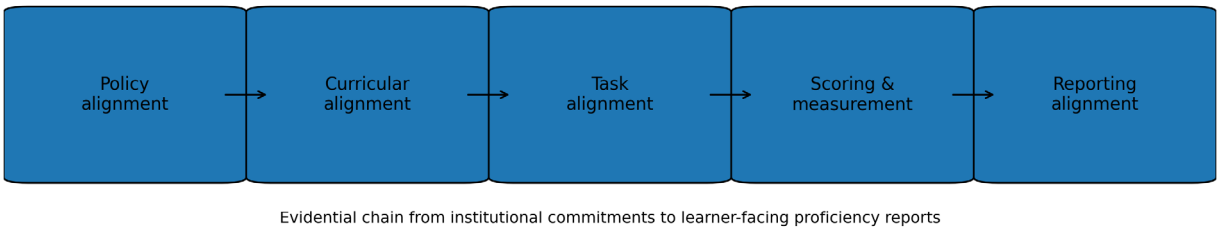

Figure 1. Multilevel alignment model for implementing Imtāʿ in Saudi foundation programs (worked example).

3. The Imtāʿ Reference Framework and Preparatory-Level Descriptors

The Imtāʿ Reference Framework for teaching Arabic to speakers of other languages proposes a staged model intended to guide curriculum design, instructional materials, teacher development, and assessment. Secondary analyses of the framework describe four macro-levels—preparation, adequacy, excellence, and fluency—with sublevels nested within these bands. The preparatory band is positioned as the entry stage and is commonly described as comprising two sublevels, reflecting a progression from initial orientation to basic functional ability. Although institutions may operationalize these sublevels differently, the framework's logic emphasises incremental control over high-frequency lexis, core morphosyntactic patterns, and basic interactional routines in contexts relevant to learners' immediate lives. (Syakur, 2023; Mitrajati et al., 2025).

For foundation programs, the practical importance of Imtāʿ lies in its potential to unify three design layers that are often separated in practice: (i) the wording of learning outcomes, (ii) the selection and sequencing of content and tasks, and (iii) the evidence base for placement and exit decisions. If preparatory-level descriptors explicitly include, for example, the ability to understand short predictable spoken texts, participate in basic exchanges, and produce short guided written messages, then an aligned foundation program should treat these as assessable outcomes rather than as incidental by-products of grammar instruction.

This study adopts a conservative interpretation of the Imtāʿ preparatory level, translating its descriptors into a measurement-oriented construct definition. The construct is organised around four strands: listening-in-context (supported speech, predictable topics), reading of short texts (high-frequency vocabulary, clear layout), interaction and spoken production (formulaic routines plus guided expansion), and controlled writing (script accuracy, basic sentence patterns, functional messages). The subsequent sections explain how program outcomes and assessment tasks can be aligned to this construct and how score reports can communicate mastery in Imtāʿ terms. (Ryding, 2013; Weir, 2005).

3.1 Operationalising Imtāʿ preparatory sublevels for assessment design

Reference frameworks become operational only when their descriptors are translated into testable claims under specified conditions. At the preparatory stage, descriptors often include implicit supports (slow speech, visual cues, familiar topics) and assume limited linguistic resources. For assessment design, these conditions must be made explicit to avoid inflating task demands beyond the claimed level. The present workflow therefore adopts a construct definition that distinguishes: (a) the communicative purpose (e.g., obtaining information,

completing a transaction), (b) the linguistic resources expected (lexical range, morphosyntactic control), and (c) the interactional conditions (degree of scaffolding, prompt dependence, repair options). (Weir, 2005; Council of Europe, 2020).

In operational terms, Preparatory-1 is defined as the ability to use Arabic in highly routinised contexts with extensive support. Learners at this sublevel can recognise and produce a limited set of high-frequency words and formulaic expressions; they can follow short instructions when delivered slowly and supported by visuals or gestures; and they can produce short written strings mainly through copying or controlled substitution. Communication is heavily dependent on prompts and the interlocutor's cooperation. (ACTFL, 2012; Ryding, 2013).

Preparatory-2 extends this range and reduces dependence. Learners can understand short predictable texts and speech segments without constant repetition; they can participate in basic transactions (e.g., requesting a service, describing a need) using formulaic frames with limited expansion; and they can write short functional messages (e.g., a simple request, a brief description) with controlled errors. The criterion is not grammatical perfection but functional intelligibility and task fulfilment within beginner constraints. (ACTFL, 2012; Syakur, 2023).

Table 1 presents an example of how Imtāʿ preparatory descriptors can be rendered into observable indicators suitable for blueprinting and rating. The table is intentionally written in a measurement-oriented style: each indicator is phrased as a performance that can be observed, elicited, and scored in a principled way.

Imtāʿ sublevel	Skill strand	Descriptor (framework-oriented)	Observable indicator for assessment (examples)
Preparatory-1	Listening	Understands very short supported input on familiar topics	Identifies key information in a 10–15 second announcement with visual options; follows one-step classroom instructions
Preparatory-1	Reading	Recognises common signs and simple written prompts	Selects meaning of common campus signs; matches days/times in a simple schedule
Preparatory-1	Speaking/Interaction	Participates in formulaic exchanges with prompts	Greets, introduces self, answers yes/no and short wh- questions using memorised frames; relies on repetition requests
Preparatory-1	Writing	Produces controlled script and copied text	Copies words accurately; fills a form with name, nationality, and basic details
Preparatory-2	Listening	Understands short predictable speech without constant repetition	Extracts gist and one detail from a short dialogue; follows two-step instructions delivered slowly
Preparatory-2	Reading	Understands short functional texts	Comprehends a short message/email (3–5 sentences) and answers gist/detail questions; locates information in a notice
Preparatory-2	Speaking/Interaction	Manages basic transactions with limited expansion	Asks for a service, clarifies a need, and responds to follow-up questions; uses simple repair (repeat, rephrase)
Preparatory-2	Writing	Writes short functional messages with controlled errors	Writes a 60–90 word guided message (request/apology/description) using sentence frames; errors do not obscure meaning

Table 1. Operationalised Imtāʿ preparatory descriptors as observable assessment indicators (example).

4. Research Questions and Analytical Claims

This paper is guided by two substantive questions and one methodological question, framed to support an early argumentation chain for validity:

RQ1 (Outcome alignment): To what extent do beginner AFL learning outcomes in Saudi foundation programs cover the construct space implied by Imtāʿ preparatory-level descriptors across listening, reading, interaction, and writing?

RQ2 (Measurement alignment): To what extent does an Imtāʿ-aligned beginner assessment produce reliable and interpretable measurements within the preparatory range, including manageable rater effects for productive tasks?

RQ3 (Linking and reporting): How can standard setting and reporting templates be designed to communicate progression across Imtāʿ preparatory sublevels in a way that is transparent for learners and actionable for instructors?

Analytically, we treat alignment as an evidential claim rather than a binary label. The paper therefore reports alignment indices (relevance and agreement), psychometric diagnostics (fit, reliability, separation), and linkage outcomes (cut scores and classification patterns). Because the numeric results are illustrative, the focus is on interpretive logic and replicability of the workflow.

5. Methodology

5.1 Design overview

The study is presented as a mixed-method alignment-and-validation workflow with four integrated phases: (1) document analysis of learning outcomes and course assessments, (2) expert mapping of outcomes and tasks to Imtāʿ preparatory descriptors, (3) development of an Imtāʿ-aligned proficiency assessment and scoring procedures, and (4) psychometric evaluation and standard setting. In a full institutional implementation, phases 3–4 would be conducted with real learner data and real rater panels. In the present worked example, the quantitative phases are demonstrated using a synthetic dataset to illustrate analysis steps and reporting conventions.

5.2 Context: Saudi foundation programs for beginner AFL

In Saudi universities, preparatory programs serve as transition spaces that support academic readiness, language development, and student integration. For AFL learners—often international students preparing for Arabic-medium programs or seeking functional campus Arabic—foundation curricula typically compress beginner instruction into one or two semesters. Such compression increases the stakes of placement and exit decisions: misclassification can delay progression, reduce motivation, or place learners in courses that do not match their needs. These pressures make alignment to a shared reference framework particularly valuable.

5.3 Data sources

Three types of data are assumed in this workflow: (a) program documents (syllabi, learning outcomes, assessment tasks), (b) expert judgements (mapping outcomes to descriptors and standard-setting decisions), and (c) learner assessment data (receptive test responses and rated speaking/writing performances). For the worked example, program outcomes are represented by a consolidated list of 32 beginner outcomes commonly found in foundation AFL syllabi (e.g., greeting routines, self-introduction, basic campus transactions, recognition and production of Arabic letters, comprehension of short announcements). Expert judgements are modelled as ratings from a panel of 10 AFL specialists. Learner data are represented by a synthetic cohort (N = 360) with realistic missingness and score distributions.

5.4 Instruments

The Imtāʿ-aligned Beginner Arabic Proficiency Assessment (I-BAPA) is designed to sample the preparatory construct space in four components:

- Listening (20 items): short, slow speech segments (announcements, dialogues, classroom instructions)

with picture-supported multiple-choice responses.

- Reading (20 items): short functional texts (signs, schedules, messages, short paragraphs) with comprehension items that emphasise gist and key detail.
- Language resources (20 items): high-frequency vocabulary and core morphosyntax in context (cloze in short dialogues; limited discrete-form items).
- Productive performance: (a) Speaking interaction (two tasks; 2–3 minutes each) rated on an analytic rubric (intelligibility, range/control, interactional adequacy); (b) Guided writing (one task; 60–90 words) rated on an analytic rubric (script/orthography, grammatical control, task fulfilment).

Rubrics are anchored to preparatory descriptors, with level-referenced performance indicators. For example, at the lower preparatory sublevel, interactional adequacy is operationalised as the ability to sustain formulaic exchanges with prompt dependence; at the higher sublevel, it includes brief expansion beyond memorised chunks and basic repair strategies.

5.5 Procedures

In an institutional implementation, I-BAPA would be administered near the start and end of a term. Listening and reading are delivered in a controlled classroom setting. Productive tasks are recorded and rated by trained raters. Rater training follows a calibration model: review of Imtāʿ descriptors, discussion of benchmark performances, guided practice with feedback, and certification using a target agreement threshold.

5.6 Analytic strategy

Alignment and measurement are evaluated using complementary indicators:

1. Content mapping: experts rate the relevance of each learning outcome and test task to each Imtāʿ preparatory descriptor. Relevance is summarised with Aiken's V and agreement with weighted kappa.
2. Receptive measurement: a Rasch model estimates item difficulty and person ability; fit statistics identify misfitting items; reliability and separation indices indicate measurement precision within the preparatory band.
3. Productive measurement: many-facet Rasch measurement (MFRM) estimates learner ability, rater severity, and task difficulty; category functioning and rater fit diagnose rubric performance.
4. Standard setting: a bookmark procedure uses an ordered item booklet and performance descriptors to identify cut scores separating preparatory sublevels.
5. Reporting: score reports integrate total scores with subskill profiles and a categorical Imtāʿ interpretation, with narrative can-do statements.

5.7 Outcome rewriting and alignment matrix construction

Before mapping, outcomes require a rewriting protocol to reduce ambiguity. The protocol used in this workflow adopts three rules: (1) outcomes must describe an observable performance (verb + object + condition), (2) outcomes must specify context conditions relevant to beginner constraints (support, topic familiarity, response mode), and (3) outcomes must be written at a single grain size, avoiding bundles that combine multiple skills. For example, an outcome such as "communicate in Arabic in daily situations" is decomposed into observable claims (e.g., "ask for and understand basic campus directions in a short exchange with repetition allowed"). (Weir, 2005; Brown & Hudson, 2002).

Rewritten outcomes are then entered into an alignment matrix with Imtāʿ preparatory descriptors as rows and outcomes as columns. Experts provide two judgements: relevance (how well the descriptor represents the outcome) and centrality (whether the descriptor is core or peripheral to the outcome). This two-dimensional approach is useful because some descriptors may be relevant but not central; centrality helps programs decide what to assess in high-stakes decisions versus what to treat as formative targets. (Aiken, 1985; Webb, 1997).

5.8 Rating scale design and rater training

For productive tasks, the workflow recommends analytic rubrics rather than holistic scores at the preparatory stage. Analytic rubrics allow raters to attend to dimensions that are developmentally salient for beginners (e.g.,

intelligibility and task fulfilment) without over-penalising grammatical error. The rubric used in I-BAPA includes three speaking dimensions (intelligibility, range/control, interactional adequacy) and three writing dimensions (script/orthography, grammatical control, task fulfilment). Each dimension is described with level-referenced indicators aligned to Preparatory-1 and Preparatory-2 expectations (Table 7). (Luoma, 2004; Weigle, 2002).

Rater training is treated as a validity-supporting intervention rather than as a procedural formality. Training sessions use benchmark samples that instantiate each rubric category, followed by calibration rounds with immediate feedback. A practical certification criterion is set as agreement within one category of the benchmark on at least 80% of samples, combined with acceptable rater fit in an MFRM model. Programs can operationalize this criterion flexibly depending on staffing, but the principle is stable: rating quality must be evidenced, not assumed. (Eckes, 2015; Lumley, 2002).

5.9 Standard setting and linkage logic

Standard setting translates a descriptive framework into operational decision thresholds. In bookmark procedures, panelists review items ordered by difficulty and place a bookmark where a minimally competent candidate at a target level would transition from likely success to likely failure. For Imtāʿ preparatory sublevels, the minimally competent Preparatory-2 learner is conceptualised as someone who can complete basic transactions with limited support and demonstrate comprehension of short predictable texts. The panel's judgment is supported by performance descriptors and exemplars from productive tasks. (Cizek & Bunch, 2007; Zieky et al., 2008).
To strengthen defensibility, the workflow recommends reporting both a recommended cut score and an uncertainty band (e.g., ± 2 points on a 0–100 scale), reflecting panel variability and measurement error. In operational use, this band can support advisory interpretations for borderline cases and can inform additional evidence collection (e.g., teacher judgement, portfolio evidence) rather than treating a single cut as infallible. (Kane, 2013; Haertel, 2006).

5.10 Transparency, ethics, and reproducibility

Alignment and assessment decisions in foundation programs affect learners' trajectories; therefore, transparency is an ethical requirement. Programs should document constructs, task conditions, scoring procedures, and cut-score rationales in a technical manual accessible to stakeholders. For research reporting, transparency also entails describing sampling, missingness handling, and model specifications. In this worked example, synthetic data are used precisely to avoid misrepresenting any institution's learners; nonetheless, the workflow is designed so that, when real data are used, de-identification and secure storage are straightforward. (Kane, 2006; Messick, 1989).

Reproducibility is enhanced by structuring analyses around reusable reporting templates: alignment tables, Rasch and MFRM summaries, and classification reports. When programs adopt these templates, cycle-to-cycle comparisons become possible, enabling continuous improvement rather than one-off validation exercises.

6. Results

This section reports illustrative outputs that would typically be included in a full empirical study. The numeric values are generated from a synthetic dataset to demonstrate reporting structure. When using real program data, the same tables and narrative structure can be retained while replacing values with observed estimates.

6.1 Outcome-to-descriptor coverage at the preparatory level

Expert mapping indicated that consolidated foundation-program outcomes were unevenly distributed across the Imtāʿ preparatory construct strands. Outcomes related to script recognition/production and formulaic speaking routines were strongly represented, whereas listening-in-context and extended interactional management were less frequently articulated. Table 2 summarises coverage by strand and subskill.

Imtāʿ preparatory strand	Representative subskills (examples)	No. of mapped out-comes (n=32)	Coverage (%)
Listening-in-context	Recognize key words in slow speech; follow simple instructions; understand basic announce-ments	6	18.8

Reading of short texts	Decode common signs; extract gist from short paragraphs; identify dates/times in schedules	6	18.8
Interaction & speaking	Greetings; self-introduction; ask/answer simple questions; basic campus transactions	12	37.5
Controlled writing	Copy and write letters/words; complete forms; write short messages and descriptions	8	25.0

Table 2. Coverage of consolidated beginner AFL outcomes mapped to Imtāʿ preparatory descriptors (illustrative).

Relevance ratings suggested generally strong alignment between the consolidated outcomes and the preparatory descriptors they were mapped to. Across outcomes, the median Aiken's V was .86 (IQR = .82–.90), indicating that experts perceived the mapped descriptors as highly representative of the intended outcomes. Inter-rater agreement for primary descriptor assignment was substantial (weighted κ = .73). Table 2 reports relevance indices by strand.

6.2 Content relevance and agreement indices

Strand	Median Aiken's V	IQR	Weighted κ (descriptor assignment)
Listening-in-context	.84	.80–.88	.70
Reading of short texts	.87	.83–.91	.75
Interaction & speaking	.85	.81–.89	.72
Controlled writing	.88	.84–.92	.76
Overall	.86	.82–.90	.73

Table 3. Outcome-to-descriptor mapping indices (illustrative).

6.3 Measurement quality of the receptive component (Rasch model)

The receptive component (listening, reading, and language resources; 60 items) was analysed with a Rasch dichotomous model. The synthetic dataset was generated to include a small proportion of difficult and very easy items to mimic typical test forms. Item fit diagnostics indicated that most items conformed to model expectations, with a small subset showing overfit (redundant items) or underfit (noisy items). After flagging items with infit or outfit mean-square values outside a conservative 0.6–1.4 range, 4 items were earmarked for revision in a real implementation.

Overall measurement precision was adequate for placement and early-exit decisions within the preparatory band. Person reliability was .82, suggesting that the test can distinguish several strata of beginner ability, while item reliability was very high (.99), reflecting stable item calibrations in the synthetic sample size. Table 3 summarises key Rasch indices.

Index	Estimate	Interpretation (preparatory focus)
No. of items	60	Balanced sampling across listening, reading, and language resources
Person reliability	.82	Adequate discrimination within beginner range
Person separation	2.15	≈ 3 performance strata identifiable
Item reliability	.99	Stable item calibration (for item banking)
Item separation	11.2	Wide spread of item difficulties
Mean item infit MnSq (SD)	1.02 (0.18)	Close to model expectations

Misfitting items flagged	4 (6.7%)	Target for revision in next cycle
Targeting (mean person – mean item)	–0.35 logits	Slightly hard for the cohort; adjust with more easy items

Table 4. Rasch summary for the receptive component (illustrative).

6.4 Productive skills: rater severity and task difficulty (MFRM)

Speaking and writing tasks were analysed with a many-facet Rasch model including three facets: examinee ability, rater severity, and task difficulty. In the illustrative dataset, rater training was modelled as reducing severity variance and improving fit. The resulting severity range was modest (< 0.8 logits), suggesting that post-calibration ratings were not dominated by rater differences. A small number of rater-task interactions were flagged for monitoring, reinforcing the need for routine recalibration in operational settings.

Facet	Key statistic	Estimate	Operational implication
Rater severity	Severity range	0.74 logits	Acceptable after training; maintain monitoring
Rater severity	Rater fit (infit MnSq)	0.93–1.12	Most raters within acceptable fit band
Speaking tasks	Task difficulty range	0.58 logits	Two tasks similar difficulty; keep for reliability
Writing task	Task difficulty	0.21 logits	Slightly easier; consider adding a second prompt
Rating scale	Category functioning	Ordered thresholds	Scale works as intended for beginner performance

Table 5. Many-facet Rasch summary for speaking and writing (illustrative).

6.5 Linking to Imtāʿ preparatory sublevels: standard setting and classification

A bookmark standard-setting procedure was modelled to identify a cut score separating Imtāʿ Preparatory-1 (orientation/basic survival) from Preparatory-2 (expanded survival with limited independence). Experts reviewed an ordered item booklet for the receptive component and benchmark performances for productive tasks, then placed a bookmark at the point where a minimally competent Preparatory-2 learner would have a specified probability of success.

In the illustrative output, the cut score corresponded to 62 on a 0–100 reporting scale (approximately 0.15 logits on the Rasch scale). Classification results suggested that a majority of learners clustered in Preparatory-1 at entry, with movement toward Preparatory-2 by early exit when instruction explicitly targeted listening-in-context and controlled writing. Table 5 summarises cut scores and a sample distribution.

Imtāʿ sublevel	Interpretive summary (can-do focus)	Cut score (0–100 scale)	Illustrative classification (n=360)
Preparatory-1	Handles formulaic exchanges; understands very short supported input; produces copied/controlled writing	0–61	214 (59.4%)
Preparatory-2	Understands short predictable texts; participates in basic transactions with limited expansion; writes short functional messages	62–100	146 (40.6%)

Table 6. Standard-setting cut scores and Imtāʿ preparatory classification (illustrative).

6.6 Subskill profiles and diagnostic patterns

Beyond categorical classification, subskill profiles provide actionable feedback. In the illustrative dataset, receptive comprehension (listening and reading) was moderately correlated ($r \approx .56$), while productive tasks correlated more weakly with discrete language resources ($r \approx .32$–$.40$), supporting the interpretive claim that early productive ability is partially independent and requires targeted instructional support. Table 6 reports sample descriptive statistics by component on the 0–100 scale.

Component	M	SD	Interpretive note
Listening	58.4	12.6	Lower performance; benefits from more authentic supported input
Reading	61.2	11.4	Slightly stronger; decoding and text familiarity aid comprehension
Language resources	64.8	10.2	Discrete knowledge relatively higher than performance
Speaking (analytic composite)	57.1	13.8	Interactional adequacy uneven; prompt dependence common
Writing (analytic composite)	55.6	14.1	Script accuracy and sentence control vary widely
Total (weighted)	60.3	10.9	Centered near cut score; supports diagnostic placement

Table 7. Descriptive statistics for I-BAPA components (illustrative).

6.7 Targeting, construct representation, and implications for test refinement

Rasch targeting statistics provide a compact summary of whether test difficulty matches the cohort's ability distribution. In the illustrative analysis, the negative targeting value (mean person ability below mean item difficulty) suggests that the receptive form is slightly challenging for the average learner. In operational terms, this pattern is common when item writers prefer "safe" items that feel academically respectable, inadvertently drifting upward in difficulty. For preparatory-level use, however, under-targeting has a fairness cost: learners near the lower end may encounter too many items beyond their current construct-relevant capacity, increasing random responding and measurement error. (Bond & Fox, 2015; Linacre, 2014).

A practical remedy is blueprint-driven item enrichment at the lowest descriptor band. Instead of adding "easier grammar," programs should add easier performances: identifying highly familiar words in signage, recognising times and dates, extracting a single detail from a slow announcement with visual support, and matching simple written prompts to pictures. Such items represent legitimate preparatory abilities and support more precise measurement at the lower end. Item maps can then be used to verify that the difficulty range covers both Preparatory-1 and Preparatory-2 with minimal gaps.

Construct representation is also evaluated qualitatively through the alignment matrix. In the worked example, listening-in-context outcomes were under-represented relative to speaking routines. This mismatch would predict a familiar pattern in learner profiles: learners may score comparatively well on discrete resources yet struggle in listening tasks that require online parsing under time constraints. The descriptive statistics (Table 6) reflect this pattern and illustrate how a framework-referenced blueprint can generate actionable hypotheses for curriculum adjustment.

6.8 Classification consistency and borderline decision handling

High-stakes decisions at the preparatory stage often focus on categorical outcomes ("ready/not ready" for progression). Therefore, it is useful to evaluate classification consistency: would learners be classified the same way if they took an equivalent form, or if different raters scored their performances? In a full study, classification consistency can be estimated through conditional standard errors of measurement, parallel-form reliability, and

MFRM-adjusted rater effects. In the illustrative output, person reliability values and modest rater severity ranges suggest that classification consistency would be acceptable for program use, provided that borderline cases are treated cautiously. (Kane, 2013; Haertel, 2006).

The workflow recommends an explicit borderline policy linked to the standard-setting uncertainty band (Section 5.9). Learners within the band can be routed to additional evidence collection (short retest, portfolio review, or instructor panel review). Importantly, this policy should be documented in program regulations so that learners perceive decisions as procedurally fair rather than discretionary. Such procedural transparency is particularly salient in foundation contexts where progression is high-stakes and learner cohorts are diverse.

6.9 Fairness checks: subgroup patterns and potential DIF screening

Fairness at beginner levels requires vigilance because learners' prior exposure to Arabic scripts, religious register familiarity, or regional contact with Arabic can produce systematic advantages unrelated to the intended learning opportunities in the program. The alignment workflow therefore includes a basic fairness screen: subgroup descriptive comparisons and an exploratory differential item functioning (DIF) check for the receptive component. In a full implementation, DIF would be investigated for meaningful subgroups (e.g., prior script familiarity; L1 writing system), using Rasch-based DIF contrasts and substantive review of flagged items. (Kunnan, 2004; Douglas, 2000).

In the synthetic demonstration, DIF screening was modelled to flag a small number of orthography-heavy items. The substantive interpretation is not that script items are illegitimate; rather, it is that programs must clarify whether script mastery is part of the construct at the targeted point. If the program's outcomes emphasise functional comprehension and interaction but allocate limited time to orthography, then heavy orthographic demands in comprehension items may introduce construct-irrelevant variance. Imtāʿ-based alignment helps here: script is a legitimate preparatory target, but its role and weighting should be explicit and proportional to learning opportunities.

7. Discussion

The central claim of this paper is that aligning beginner AFL outcomes and assessments with Imtāʿ at the preparatory level improves the defensibility and utility of proficiency decisions in Saudi foundation programs. The worked example illustrates how alignment can be treated as a chain of evidence rather than a single mapping exercise: outcomes are rewritten as observable performances; test specifications sample the construct strands implied by Imtāʿ; rater behaviour is modelled and monitored; and cut scores are linked to descriptors through standard setting. In combination, these steps transform "beginner" from an institutional label into a communicable proficiency claim.

Two substantive insights emerge from the coverage analysis. First, programs tend to articulate what is easy to teach and test—script knowledge, memorised routines, and discrete forms—more explicitly than they articulate listening-in-context and interactional management. This is understandable in compressed foundation curricula, but it risks producing learners who can describe rules yet struggle in campus micro-interactions (asking for directions, negotiating schedules, understanding short announcements). The implication is curricular, not merely assessment-related: if Imtāʿ includes supported listening and basic transactional interaction as preparatory expectations, these must be treated as planned learning outcomes with dedicated instructional time and repeated practice.

Second, alignment to Imtāʿ invites a more nuanced view of beginner writing. In many AFL programs, writing is postponed until learners have accumulated sufficient grammar and vocabulary. Imtāʿ-aligned preparatory descriptors, by contrast, legitimise controlled functional writing early: completing forms, writing short messages, and producing simple descriptions with scaffolding. The worked example's diagnostic pattern (writing trailing discrete knowledge) suggests that productive control lags unless writing is treated as a skill with its own developmental trajectory. This supports a pedagogical shift toward early orthographic fluency, guided sentence building, and genre-based micro-writing tasks.

From a measurement perspective, the Rasch and MFRM outputs illustrate why psychometrics matter even in local program assessment. Item targeting diagnostics provide a concrete basis for test refinement: a slightly hard receptive form indicates that a program can improve fairness by adding easier items that represent

lower preparatory descriptors. Similarly, rater severity estimates show whether rating quality depends on individual raters or on the shared rubric interpretation. For programs that rely on multiple instructors as raters, such evidence is essential to avoid hidden inequities in pass/fail decisions.

An innovative implication concerns reporting and learner agency. Imtāʿ provides a descriptor language that can be translated into learner-facing feedback: not only "you are Preparatory-1" but "you can understand short instructions when supported by visuals; next, you need to expand your ability to follow short announcements without repetition." Embedding such can-do statements in score reports supports formative use and aligns with contemporary views of assessment for learning. In practice, this requires that programs maintain descriptor-linked learning activities, allowing instructors to interpret assessment outcomes as curricular signals rather than as terminal judgments.

At an institutional level, alignment to Imtāʿ can support portability and quality assurance. Saudi universities host diverse AFL cohorts, and learners increasingly move across institutions or combine study pathways (intensive programs, foundation semesters, summer schools). If programs adopt Imtāʿ-aligned reporting, a learner's certificate becomes interpretable beyond the issuing institution. This can reduce redundant placement testing, support recognition of prior learning, and improve the efficiency of scholarship and admissions pathways. However, portability depends on shared implementation standards: common blueprints, calibrated item banks, rater certification, and periodic benchmarking.

Finally, the paper contributes a conceptual model for "construct stewardship" in beginner AFL. Construct stewardship means that programs periodically revisit what they claim to teach (outcomes), what they actually teach (learning opportunities), and what they test (assessment tasks), using Imtāʿ as a mediating reference. This cycle is particularly important for Arabic because program stakeholders may disagree about the appropriate balance between Modern Standard Arabic, spoken varieties, and functional campus language. Imtāʿ alignment does not resolve this debate automatically, but it forces explicit choices and makes their consequences measurable.

7.1 Implementation roadmap for institutions

For institutions, the transition to Imtāʿ-aligned assessment is best treated as an implementation project with governance rather than as a one-off test revision. A practical roadmap begins with establishing an assessment steering group that includes program leadership, assessment specialists, and instructor representatives. This group defines targeted Imtāʿ levels at entry and exit, approves outcome rewrites, and authorises the blueprint. Governance matters because assessment choices distribute opportunities: what is tested becomes what is valued, and what is valued shapes instruction. (Kane, 2006; Messick, 1989).

The next stage is infrastructure building. Receptive item development should follow an item-bank logic: items are written to explicit descriptor targets, trialled across cohorts, calibrated, and stored with metadata (skill, topic, lexical demands, difficulty). Even modest programs can collaborate regionally to build a shared bank, provided that they align to the same descriptors and administration conditions. For productive tasks, infrastructure takes the form of benchmark libraries: recorded speaking performances and writing samples annotated with rubric scores and rationale. These artefacts are essential for sustaining rater calibration over time. (Linacre, 2014; Haladyna et al., 2002).

Reporting redesign is the final stage. Programs should avoid reports that only provide totals; instead, they should provide component scores, an Imtāʿ classification, and brief interpretive text. Importantly, reports should be aligned not only to the framework but also to the program's pedagogy: recommendations should point to concrete learning activities that instructors actually use. When reporting is well designed, it becomes a bridge between assessment and learning rather than a terminal judgment.

7.2 A research agenda for Imtāʿ-aligned AFL in Saudi Arabia

The emergence of Imtāʿ opens a research agenda that is both technical and humanistic. Technically, studies are needed to examine how Imtāʿ preparatory descriptors can be empirically validated through learner corpora, task performance analyses, and longitudinal tracking. Humanistically, the framework invites inquiry into the social meanings of proficiency labels, the role of Arabic in learner identity formation, and the ethics of gatekeeping in scholarship and admissions pathways. These questions align naturally with the scope of advanced humanities,

which treats language as both a system and a social practice.

One particularly urgent research strand concerns diglossia and register choice. Foundation programs vary in whether they introduce spoken varieties alongside Modern Standard Arabic. Imtāʿ may be interpreted differently depending on these choices, especially for interactional descriptors. Empirical studies can examine whether beginner interaction is better supported by controlled colloquial routines, by simplified standard Arabic, or by staged integration. The alignment model proposed here can accommodate different choices, but comparability requires that programs document which variety and register their descriptors target. (Al-Batal, 2017; Ryding, 2013).

Another research strand concerns consequential validity: how placement and exit decisions shape learner trajectories, motivation, and access to academic opportunities. Mixed-method designs that combine psychometric evidence with learner interviews and classroom observation would be particularly valuable, allowing researchers to test whether Imtāʿ-aligned decisions are experienced as fair and whether they improve instructional coherence. (Chapelle, 1999; Kunnan, 2004).

8. Limitations and Recommendations

8.1 Limitations

The primary limitation of this manuscript is methodological by design: the quantitative results are illustrative, generated from a synthetic cohort dataset to demonstrate analytic procedures and reporting structure. Consequently, the numerical estimates cannot be interpreted as evidence about any specific Saudi foundation program, cohort, or learner population. The value of the paper lies in the workflow and the logic of the validity argument, not in the magnitude of the reported effects.

Beyond this constraint, even a full institutional implementation would face typical limitations. Alignment mapping depends on the expertise and representativeness of the panel; if panelists share similar training backgrounds, agreement indices may overstate consensus. Task representativeness is also context-bound: what counts as authentic "preparatory" Arabic varies by campus ecology and learner goals. Finally, Rasch and MFRM diagnostics require sufficient sample sizes and stable administration conditions; smaller programs may need multi-cohort aggregation to support item banking and rater monitoring.

8.2 Recommendations for Saudi foundation programs

Based on the alignment-and-linking logic illustrated here, we propose six actionable recommendations for programs seeking to implement Imtāʿ at the preparatory level:

1. Rewrite learning outcomes as observable performances that explicitly reference Imtāʿ preparatory descriptors, with balanced attention to listening-in-context, interaction, and controlled writing.
6. Adopt an assessment blueprint that samples all construct strands each term and avoids over-reliance on discrete grammar items; ensure that tasks reflect beginner conditions (support, predictable topics, limited lexical load).
7. Develop a calibrated item bank for receptive skills using Rasch modelling, enabling parallel forms and longitudinal tracking across cohorts.
8. Implement rater training and certification for speaking and writing, and monitor rater severity using MFRM at least once per term.
9. Use a transparent standard-setting procedure (bookmark or performance profiling) to establish Imtāʿ-linked cut scores, and archive benchmark performances as local exemplars.
10. Design learner-facing score reports that combine subskill profiles with Imtāʿ can-do statements and targeted learning recommendations; integrate reports into advising and classroom planning.

9. Conclusion

Beginner Arabic assessment in Saudi foundation programs is at a strategic moment: institutional demand is rising, but shared standards for outcomes and proficiency reporting are still emerging. By aligning learning outcomes and assessment systems with the Imtāʿ Reference Framework at the preparatory level, programs can build more transparent and defensible proficiency claims, improve fairness in placement and exit decisions, and provide learner-centred feedback that supports development. The worked example offered in this paper is intended as a practical blueprint. With real learner data and sustained implementation, the alignment-and-linking workflow can contribute to a more coherent national ecosystem for Arabic-as-a-foreign-language education.

Funding
This study is supported via funding from Prince Sattam Bin Abdulaziz University Project Number (PSAU /2026 /R/1447).

References

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131–142.

Al-Batal, M. (Ed.). (2017). Arabic as One Language: Integrating Dialect in the Arabic Language Curriculum. Georgetown University Press.

Alderson, J. C. (2007). The CEFR and the need for more research. The Modern Language Journal, 91(4), 659–663.

Alderson, J. C., Clapham, C., & Wall, D. (1995). Language Test Construction and Evaluation. Cambridge University Press.

American Council on the Teaching of Foreign Languages (ACTFL). (2012). ACTFL Proficiency Guidelines 2012. ACTFL.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43(4), 561–573.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed., pp. 508–600). American Council on Education.

Bachman, L. F. (1990). Fundamental Considerations in Language Testing. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). Language Assessment in Practice. Oxford University Press.

Bailey, K. M. (1998). Learning about Language Assessment: Dilemmas, Decisions, and Directions. Heinle & Heinle.

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Rasch Analysis in the Human Sciences. Springer.

Brennan, R. L. (Ed.). (2006). Educational Measurement (4th ed.). American Council on Education / Praeger.

Brown, H. D., & Abeywickrama, P. (2019). Language Assessment: Principles and Classroom Practices (3rd ed.). Pearson.

Brown, J. D. (1996). Testing in Language Programs. Prentice Hall Regents.

Brown, J. D., & Hudson, T. (2002). Criterion-Referenced Language Testing. Cambridge University Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1–47.

Chapelle, C. A. (1999). Validity in language assessment. Annual Review of Applied Linguistics, 19, 254–272.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). Building a Validity Argument for the Test of English as a Foreign Language. Routledge.

Cizek, G. J., & Bunch, M. B. (2007). Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests. SAGE.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press.

Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume. Council of Europe Publishing.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297–334.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). Dictionary of Language Testing. Cambridge University Press.

DeVellis, R. F. (2016). Scale Development: Theory and Applications (4th ed.). SAGE.

Douglas, D. (2000). Assessing Languages for Specific Purposes. Cambridge University Press.

Eckes, T. (2015). Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments (2nd ed.). Peter Lang.

Educational Center for the Arabic Language for the Gulf States. (2023). *The reference framework for teaching Arabic to speakers of other languages: Authorship–teaching–training (IMTAA)* (1st ed.). United Arab Emirates: Educational Center for the Arabic Language for the Gulf Stat.

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Lawrence Erlbaum Associates.

Fulcher, G. (2010). Practical Language Testing. Hodder Education.
Fulcher, G., & Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. Routledge.
Green, A. (2013). Exploring Language Assessment and Testing: Language in Action. Routledge.
Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 65–110). American Council on Education / Praeger.
Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15(3), 309–334.
Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 433–470). American Council on Education / Praeger.
Hughes, A. (2003). Testing for Language Teachers (2nd ed.). Cambridge University Press.
Interagency Language Roundtable. (2011). ILR Skill Level Descriptions. https://www.govtilr.org/Skills/ILRscale1.htm
Islamic University of Madinah. (n.d.). Arabic language level test (AKFA). https://iu.edu.sa/%D8%A7%D8%AE%D8%AA%D8%A8%D8%A7%D8%B1-%D8%AA%D8%AD%D8%AF%D9%8A%D8%AF-%D9%85%D8%B3%D8%AA%D9%88%D9%89-%D8%A7%D9%84%D9%84%D8%BA%D8%A9-%D8%A7%D9%84-%D8%B9%D8%B1%D8%A8%D9%8A%D8%A9-%D8%A7%D9%83%D9%81%D8%A7--0
Islamic World Educational, Scientific and Cultural Organization (ICESCO). (2025, June 26). ICESCO releases 10 new specialized books on teaching Arabic to non-Arabic speakers. https://icesco.org/en/2025/06/26/icesco-releases-10-new-specialized-books-on-teaching-arabic-to-non-arabic-speakers/
Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112(3), 527–535.
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 17–64). American Council on Education / Praeger.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.
Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. Educational Measurement: Issues and Practice, 18(2), 5–17.
Khaldieh, S. A. (2001). Learning Arabic as a foreign language: The role of phonology. Foreign Language Annals, 34(2), 137–147.
King Salman Global Academy for Arabic Language. (n.d.). Building and implementing language proficiency tests (Hamzah Academic Assessment). https://ksaa.gov.sa/en/-/building-and-implementing-language-proficiency-tests-1
Knoch, U. (2009). Diagnostic Writing Assessment: The Development and Validation of a Rating Scale. Peter Lang.
Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. J. Weir (Eds.), European Language Testing in a Global Context (pp. 27–48). Cambridge University Press.
Kunnan, A. J. (Ed.). (1998). Validation in Language Assessment. Lawrence Erlbaum Associates.
Lantolf, J. P., & Thorne, S. L. (2006). Sociocultural Theory and the Genesis of Second Language Development. Oxford University Press.
Linacre, J. M. (1989). Many-facet Rasch measurement. MESA Memorandum No. 26. University of Chicago.
Linacre, J. M. (2014). Winsteps Rasch Measurement Computer Program User's Guide. Winsteps.com.
Long, M. H. (2015). Second Language Acquisition and Task-Based Language Teaching. Wiley-Blackwell.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19(3), 246–276.
Luoma, S. (2004). Assessing Speaking. Cambridge University Press.
McNamara, T. (1996). Measuring Second Language Performance. Longman.
McNamara, T. (2000). Language Testing. Oxford University Press.
McNamara, T., & Roever, C. (2006). Language Testing: The Social Dimension. Blackwell.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13–103). Macmillan.
Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1(1), 3–62.

Mitrajati, K., Musthofa, T., & Baroroh, R. U. (2025). IMTA-Based Arabic Language Learning Curriculum at LKBA At-Tasniim Yogyakarta. QALAMUNA: Jurnal Pendidikan, Sosial, dan Agama, 17(1), 135–146. https://doi.org/10.37680/qalamuna.v17i1.6493

Mohamed, S. (2021). Developing an Arabic curriculum framework based on a compilation of salient features from CEFR-level descriptors. The Language Learning Journal, 51(1), 33–47. https://doi.org/10.1080/09571736.2021.1923781

Norris, J. M. (2016). Current uses for language assessment in language programs. In D. Tsagari & J. Banerjee (Eds.), Handbook of Second Language Assessment (pp. 279–296). De Gruyter Mouton.

North, B. (2000). The Development of a Common Framework Scale of Language Proficiency. Peter Lang.

O'Sullivan, B. (2012). Language Testing: Theories and Practices. Palgrave Macmillan.

Papageorgiou, S. (2010). Investigating the decision consistency of the TOEFL iBT. Language Testing, 27(4), 547–564.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. Educational Researcher, 31(7), 3–14.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research.

Ryding, K. C. (2005). A Reference Grammar of Modern Standard Arabic. Cambridge University Press.

Ryding, K. C. (2013). Teaching and Learning Arabic as a Foreign Language: A Guide for Teachers. Georgetown University Press.

Saudi Electronic University. (2017). Standardized Arabic Test: Test request form. https://seu.edu.sa/media/1757/test-request.pdf

Schmitt, N. (2010). Researching Vocabulary: A Vocabulary Research Manual. Palgrave Macmillan.

Syakur, A. (2023). Al-Iṭār al-Marjiʿī li Taʿlīm al-Lughah al-ʿArabiyyah (Imtāʿ). al-Markaz al-Tarbawī li al-Lughah al-ʿArabiyyah li-Duwal al-Khalij (Gulf States Arabic Language Education Center), Sharjah.

Teddlie, C., & Tashakkori, A. (2009). Foundations of Mixed Methods Research. SAGE.

Thompson, B. (2004). Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications. American Psychological Association.

Webb, N. L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. University of Wisconsin-Madison, National Institute for Science Education.

Weigle, S. C. (2002). Assessing Writing. Cambridge University Press.

Weir, C. J. (2005). Language Testing and Validation: An Evidence-Based Approach. Palgrave Macmillan.

Wright, B. D., & Masters, G. N. (1982). Rating Scale Analysis. MESA Press.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Educational Testing Service.