International Journal of Arabic Language Teaching
P-ISSN: 2684-690X E-ISSN: 2686-214X
https://e-journal.metrouniv.ac.id/index.php/IJALT

Development and Preliminary Validation of a Two-Tier Diagnostic Test for Identifying Misconceptions in Arabic Grammar Learning

Siti Zulfa Hidayatul Maula[1]*, Mohammad Ahsanuddin[2], Laily Maziyah[3], Moh. Zawawi[4], Muhammad Bahrul Ula[5]
[1,2,3]Universitas Negeri Malang, Indonesia
[4]Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia
[5]Universitas Al-Falah As-Sunniyah, Indonesia

Article History:
Received : 04 November 2025
Revised : 27 June 2026
Published : 30 June 2026

Keywords:
Arabic Grammar; Students' Misconception; Two-Tier Diagnostic Test.

***Correspondence Address:**
szulfa.hm@gmail.com

©authors 2026
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

DOI: 10.32332/ijalt.v8i01.11810

***Abstract*:** This study aims to develop and preliminarily validate a two-tier diagnostic test instrument to detect students' misconceptions on *fi'l*, *adawāt al-istifhām*, and *ẓarf* in the context of Arabic language learning at Madrasah Aliyah. Using the Plomp model, the researchers developed 10 items administered to 10th grade students through a small-scale trial (n=9) and a large-scale trial (n=31), both conducted within a single class at one school. The findings were as follows; First, construct validation yielded a mean score of 4.21 (84.2%) and content validation yielded 4.42 (88%), both meeting the very valid criterion. Second, item analysis in the large-scale trial produced correlation coefficients ranging from 0.41 to 0.83, with all items declared valid (r-calculated > r-table 0.3009). The instrument demonstrated good internal consistency (Cronbach's α = .86), with 80% of items classified as moderate in difficulty and 80% of items demonstrating good discriminating power. Third, the two-tier diagnostic test indicated that 48% of students fell into the Sound Understanding category, 10% into Lucky Guess, 8% into Misconception, and 33% into Lack of Knowledge. These findings suggest that the instrument shows preliminary adequacy for classifying student understanding in Arabic grammar, though further validation with larger and more diverse samples is recommended.

INTRODUCTION

Assessment plays a vital role in education by measuring students' capabilities and identifying learning obstacles.[1] Among various evaluation tools available, the two-tier diagnostic test has received considerable attention as an instrument for detecting student misconceptions.[2] Treagust defines the two-tier

[1] Irit Levy-feldman, "The Role of Assessment in Improving Education and Promoting Educational Equity," *Education Sciences* 15, no. 2 (2025): 1–11, https://doi.org/https://doi.org/10.3390/educsci15020224.

[2] Tukiyo et al., "The Development of A Two-Tier Diagnostic Test to Detect Student's Misconceptions in Learning Process," *MUDIR (Jurnal Manajemen Pendidikan)* 5, no. 1 (2023): 92–96, https://doi.org/10.55352/mudir.

diagnostic test as an instrument consisting of two levels, in which the first level contains multiple-choice questions measuring students' conceptual understanding, while the second level contains reasoning options that reflect the rationale behind the answers given in the first level.[3] It is the combination of responses across both levels that enables this instrument to classify student understanding in greater detail than conventional multiple-choice tests.[4]

Based on the combination of responses across both levels, Treagust classifies student understanding into four diagnostic categories,[5] namely Sound Understanding (SU), Lucky Guess (LG), Misconception (MC), and Lack of Knowledge (LK). Sound Understanding refers to cases where students answer correctly on both levels, while Lucky Guess describes students who answer correctly on the first level but incorrectly on the second. Misconception occurs when students answer incorrectly on the first level but provide consistent reasoning on the second, whereas Lack of Knowledge applies when students answer incorrectly on both levels. Among these four categories, MC is generally regarded as the most pedagogically important, as students in this category tend to hold consistently incorrect beliefs that may be resistant to change without targeted conceptual intervention.[6]

It is important to note that the distinction between MC and LK carries theoretical significance. Misconceptions in the strict sense refer to relatively stable alternative conceptions, meaning structured beliefs that conflict with accepted understanding, whereas LK indicates the absence of an established conceptual framework rather than the presence of an erroneous one. This distinction has implications for instructional design, as the two categories may call for different pedagogical responses.[7] Tuysuz further notes that the two-tier format helps reduce measurement error caused by guessing, given that conventional multiple-choice tests carry an estimated 20% probability of students answering correctly by

---

[3] David F Treagust, "Diagnostic Assessment in Science as A Means to Improving Teaching, Learning and Retention," in *UniServe Science Assessment Symposium Proceedings* (Proceedings of The Australian Conference on Science and Mathematics Education, 2012), 1–9.

[4] Handjoko Permana et al., "Development of a Two-Tier Diagnostic Test for The Analysis of Momentum and Impulse Learning Difficulties in Students," *Jurnal Penelitian & Pengembangan Pendidikan Fisika* 9, no. 2 (December 31, 2023): 237–44, https://doi.org/10.21009/1.09206.

[5] Rifat Shafwatul Anam and Surya Gumilar, "A Systematic Literature Review of Tier Tests as Diagnostic Tools in Specific Areas of Science," *Tadris: Jurnal Keguruan Dan Ilmu Tarbiyah* 9, no. 1 (June 29, 2024): 255–70, https://doi.org/10.24042/tadris.v9i1.17717.

[6] Derya Kaltakci Gurel, Ali Eryılmaz, and Lillian Christie Mcdermott, "A Review and Comparison of Diagnostic Instruments to Identify Students ' Misconceptions in Science," *Eurasia Journal of Mathematics, Science & Technology Education* 11, no. 5 (2015): 989–1008, https://doi.org/10.12973/eurasia.2015.1369a.

[7] Dyah Aniza Kismiati and Leonard R Hutasoit, "Teacher Misconceptions : A Phenomenon of the Lack of Knowledge in Science Subjects," *JPPIPA: Jurnal Penelitian Pendidikan IPA* 10, no. 14 (2024): 3493–3500, https://doi.org/10.29303/jppipa.v10i6.7226.

chance, whereas the two-tier diagnostic test is estimated to reduce this probability to around 4%.[8]

Several studies have been conducted on the development of two-tier diagnostic tests across various contexts, covering topics such as basic English grammar[9], acids and bases[10], and optical instruments[11]. The majority of these studies, however, have focused on science subjects, reflecting a broader pattern in which the two-tier framework has been more extensively developed within science education than within language education. Its application to Arabic grammar in particular has received little attention in existing literature. Although at least one study has extended the framework to English grammar, the morphosyntactic complexity of Arabic and the central role of *iʿrāb* differ substantially from English, suggesting that adapting the instrument to Arabic requires not merely a change of subject matter but a reconsideration of how its diagnostic categories are operationalized in light of the specific error patterns Arabic learners are likely to exhibit.[12]

The complexity of Arabic lies in the inseparable interconnection between syntax, morphology, and semantics,[13] such that students' errors in applying *iʿrāb* or *sharf* tend to reflect interrelated difficulties spanning the entire linguistic system rather than isolated conceptual errors. This interconnectedness makes the two-tier format particularly relevant for the Arabic language context. The first-tier measures conceptual understanding, while the second tier reveals whether students' responses reflect accurate understanding, erroneous beliefs, or mere guessing.

The instrument in this study was developed drawing on three theoretical foundations. The first is Treagust's two-tier diagnostic framework, which provides the structural basis for combining a conceptual question in the first tier with a reasoning question in the second tier. The second is error analysis theory, which holds that learner errors tend to be systematic and traceable to identifiable sources

---

[8] Gilang Ramadhan, Pratiwi Dwijananti, and Siti Wahyuni, "Analisis Kemampuan Berpikir Tingkat Tinggi (High Order Thinking Skills) Menggunakan Instrumen Two Tier Multiple Choice Materi Konsep Dan Fenomena Kuantum Siswa SMA Di Kabupaten Cilacap," *UPEJ: Unnes Physics Education Journal* 7, no. 3 (2018): 85–90, http://journal.unnes.ac.id/sju/index.php/upej.

[9] Kartika Eva Rahmawati, "Using Iconicity in Two-Tier Grammar Test for Adult Learners: A Cognitive Linguistic Approach," *ELECTRUM (The English Language and Education Spectrum)* 2, no. 2 (July 2022): 1–7.

[10] Nur Laeli Azizah, Lina Mahardiani, and Sri Yamtinah, "Analisis Miskonsepsi Dengan Tes Diagnostik Two-Tier Multiple Choice Dan In-Depth Interview Pada Materi Asam Basa," *Jurnal Pendidikan Kimia* 11, no. 2 (2022): 168–77, https://doi.org/10.20961/jpkim.v11i2.60345.

[11] Deka Maulidiansyah, Inarsih Meutia, and Edi Istiyono, "Computer-Based Two-Tier Diagnostic Test to Identify Critical Thinking Skills in Optical Instrument," in *Proceedings of the 6th International Seminar on Science Education (ISSE 2020)*, vol. 541, 2021, 413–18, https://doi.org/10.2991/assehr.k.210326.059.

[12] Neimat Idris Moh. Saeed Omer, "Error Analysis of Grammatical Features in Saudi Undergraduate EFL Writing," *ESIC* 8, no. 1 (2024): 483–98.

[13] Ameerah Alothman and Abdulmalik Alsalman, "An Arabic Grammar Auditor Based on Dependency Grammar," *Hindawi: Advances in Human-Computer Interaction* 2020, no. 1 (2020): 1–10, https://doi.org/https://doi.org/10.1155/2020/8856843.

such as first language interference and overgeneralization of target language rules.[14] The third is conceptual change theory, which suggests that some erroneous understandings may persist because they are rooted in prior knowledge structures that conflict with new material.

Dulay, Burt, and Krashen argue that errors in foreign language learning generally occur due to mother tongue transfer and the tendency to overgeneralize target language rules.[15] In the *fiʿl* material, students tend to struggle with distinguishing *fiʿl māḍī, muḍāriʿ*, and *amr*, as Indonesian does not recognize verb form changes based on tense. In the *adawāt al-istifhām* material, students often select inappropriate question words, as Indonesian question words tend to function more broadly than their Arabic counterparts. In the *zharf* material, students tend to conflate *zharf zamān* and *zharf makān* due to unfamiliarity with *iʿrāb*-governed adverbial rules. In addition, other factors such as instructional quality, limited exposure, and memorization-based learning may also play a role in this pattern, although their relative contribution remains difficult to ascertain without further investigation. Such variations can be difficult to identify through conventional multiple-choice tests, as an incorrect answer alone does not necessarily indicate whether it stems from a genuine misconception, a lucky guess, or a simple lack of knowledge, each of which may call for a different instructional response.[16]

Arabic grammar learning at the Madrasah Aliyah level encompasses more than 20 grammatical aspects distributed progressively across grades 10 through 12. Based on evaluation results at MA Ma'arif NU Kencong in the 2024/2025 academic year, the average Arabic language score for 10th grade students was 64.07, with 55.81% falling below the Minimum Completion Criteria (KKM) of 70. The assessment used was conventional in nature, limited to indicating correct and incorrect answers without revealing underlying patterns of student understanding, leaving teachers with limited means to identify the specific nature of each student's difficulties.[17] This is particularly relevant given that Arabic grammar constitutes a foundational aspect of language competence, enabling students to construct sentences, analyze texts, and access meanings in religious

[14] Mamnunah, Mamnunah, Maman Abdurrahman, and Asep Sopian, "The Error Analysis of Arabic Grammar in The Kalamuna Book," *Arabi: Journal of Arabic Studies* 6, no. 2 (2021): 158–66, https://doi.org/http://dx.doi.org/10.24865/ajas.v6i2.351.

[15] Ilham Fatkhu Romadhon, "Analisis Kesalahan Berbahasa Arab Pada Ketrampilan Berbicara Dalam Debat Bahasa Arab Di Indonesia," *Indonesian Journal of Arabic Education and Learning* 1, no. 1 (2025): 35–55, https://doi.org/10.58363/ijael.

[16] Bashar Alhafni et al., "Advancements in Arabic Grammatical Error Detection and Correction : An Empirical Investigation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, 2023), 6430–48, https://doi.org/https://doi.org/10.18653/v1/2023.emnlp-main.396.

[17] Wan Ab Aziz Wan Daud et al., "Exploring the Challenges and Strategies of Learning Arabic Language among Primary School Teachers," *Forum for Linguistic Studies* 7, no. 2 (2025): 94–103, https://doi.org/10.30564/fls.v7i2.8002.

and classical texts, as well as supporting those pursuing further study in fields requiring strong Arabic language competence.[18]

Therefore, this study aims to develop and preliminarily validate a two-tier diagnostic test specifically designed for Arabic grammar assessment at the Madrasah Aliyah level, taking into account the linguistic complexity that distinguishes this context from previous applications of the instrument. By adapting Treagust's framework to the materials of *fiʿl, adawāt al-istifhām*, and *ẓarf*, this study seeks to produce an instrument that can potentially differentiate students across the SU, LG, MC, and LK categories, and in doing so, provide preliminary diagnostic information that may support teachers in designing more targeted instructional interventions.

METHOD

This study employed a research and development (R&D) design. Sugiyono states that this type of research is used to produce a specific product and to test its effectiveness. The development model used in this study was the Plomp model. The Plomp model is also referred to as design research. Plomp states that this model consists of three stages, namely preliminary research, prototyping stage, and assessment phase. The research was conducted at MA Ma'arif NU Kencong during the 2024/2025 academic year. The subjects of this study were two expert validators, namely a lecturer in Arabic Language and Literature at UIN Maulana Malik Ibrahim Malang and a teacher at Madrasah Aliyah Assunniyyah Al-Jauhari Kencong. The object of this study was a set of two-tier diagnostic test instruments covering the materials of *fiʿl, adawāt al-istifhām,* and *zharf* for 10th grade students at MA Ma'arif NU Kencong.

Data was collected through interviews, classroom observations, validation questionnaires, and tests. The instruments used in data collection included interview guidelines, validation questionnaire sheets, and the test instrument itself. The validation questionnaire used a five-point Likert scale. Information regarding the eligibility criteria can be seen below:

Table 1. Eligibility Criteria for Validation Scores

Value	Score Range	Percentage	Criteria	Description
5	4,21 – 5,00	84% - 100%	Very Valid	Very suitable to use
4	3,41 – 4,20	68% - 84%	Valid	Suitable to use
3	2,61 – 3,40	52% - 68%	Fairly Valid	Fairly suitable to use
2	1,81 – 2,60	36% - 52%	Invalid	Not suitable to use
1	1,00 – 1,80	20% - 36%	Very Invalid	highly unsuitable for use

[18] Muhammad Rifqi Ridho and Abu Bakar, "Signifikansi Gramatika Arab Dalam Menginterpretasikan Al-Qur'an: Tinjauan Ilmu Nahwu, Ilmu Sharaf, Dan Ilmu Balaghah," *AL-MAZAYA Jurnal Pendidikan Bahasa Arab* 3, no. 1 (2025): 30–44.

Content validation assessed material suitability, language and writing quality, and relevance to the curriculum. Construct validation assessed the two-tier diagnostic test structure, distractor quality, and alignment with two-tier theory. To assess the level of agreement between the two validators beyond percentage scores alone, inter-rater reliability was calculated using Gwet's AC1 coefficient.

After expert validation, the instrument was administered to students for item analysis, covering validity, reliability, difficulty level, and discriminating power. Item validity was calculated using the Pearson Product-Moment Correlation formula, with items declared valid if r-calculated exceeded r-table at a significance level of 5% (df = N-2). Reliability was assessed using Cronbach's Alpha, with a coefficient above 0.6 indicating adequate reliability. Difficulty level was classified as easy ($P > 0.70$), moderate ($0.30 \leq P \leq 0.70$), or difficult ($P < 0.30$). Discriminating power was classified as poor ($DP \leq 0.20$), moderate ($0.20 < DP \leq 0.40$), good ($0.40 < DP \leq 0.70$), or very good ($DP > 0.70$), with items classified as moderate, good, or very good considered to have effective discriminating power.

Two-tier diagnostic test analysis was conducted to classify student responses into four diagnostic categories following Treagust's framework. Sound Understanding (SU) refers to responses that are correct on both levels, Lucky Guess (LG) refers to responses that are correct on the first level but incorrect on the second, Misconception (MC) refers to responses that are incorrect on the first level but accompanied by consistent reasoning on the second, and Lack of Knowledge (LK) refers to responses that are incorrect on both levels.

The instrument was administered in two phases following expert validation. A small-scale trial was first conducted with 9 purposively selected students, composed of 3 high-ability, 3 medium-ability, and 3 low-ability students, to obtain preliminary feedback on item quality prior to broader implementation. A large-scale trial was subsequently conducted with the full class population of 31 students. Both trials were conducted within a single class at one school. The limitation of this sample size is acknowledged, as the primary focus of the current phase was to examine the instrument's ability to classify students into the four diagnostic categories rather than to establish broad psychometric generalizability. Further research involving larger and more diverse samples from multiple schools is therefore recommended to more thoroughly verify the instrument's psychometric stability.

RESULT AND DISCUSSION

The development of the two-tier diagnostic test instrument to detect students' misconceptions on the materials of *fiʿl, adawāt al-istifhām*, and *ẓharf* consisted of three stages following the Plomp model, namely preliminary research, prototyping stage, and assessment phase.

Preliminary Research

The preliminary research stage was divided into four steps, namely needs analysis, context analysis, literature study, and conceptual framework development. The needs analysis stage was conducted through interviews with the Arabic language teacher and classroom observations at MA Ma'arif NU Kencong. The findings revealed that 55.81% of 10th grade students did not meet the Minimum Completion Criteria (KKM) with an average score of 64.07. Based on the interview results, the evaluation system in use was still conventional in nature, consisting of regular multiple-choice and essay questions that were not designed to identify the specific nature of student difficulties. The teacher also confirmed having no prior experience implementing a two-tier diagnostic test in Arabic language learning. These findings indicate a need for a more diagnostic form of assessment that can distinguish between students experiencing misconceptions, guessing, or lacking foundational knowledge.

The context analysis was conducted by examining the curriculum implemented at MA Ma'arif NU Kencong, namely the Merdeka Curriculum (Kurikulum Merdeka), with reference to the Decree of the Kemendikbudristek Number 262 of 2022 concerning the Implementation of the Merdeka Curriculum in Education Units. Under this curriculum, the grammatical materials of *fiʿl, adawāt al-istifhām,* and *ẓarf* are included in the learning scope for 10th grade students in the second semester, as referenced in the learning objectives (Capaian Pembelajaran) for Arabic language at the Madrasah Aliyah level. The learning indicators for these materials require students to identify, distinguish, and apply *fiʿl māḍī, muḍāriʿ*, and *amr* in sentence contexts, to select *adawāt al-istifhām* appropriately according to their specific semantic functions, and to identify and apply *ẓarf zamān* and *ẓarf makān* in accordance with *iʿrāb* principles.

To support the development of the two-tier diagnostic test instrument for Arabic grammar, a literature study was conducted drawing on relevant books, research journals, and other sources including online databases. The literature study covered three main areas. The first area concerned the theoretical foundations of two-tier diagnostic test development, with particular reference to Treagust's original framework and subsequent adaptations in various subject areas. The second area concerned research on misconceptions in Arabic grammar learning among Indonesian learners, focusing on documented error patterns in *fiʿl, adawāt al-istifhām*, and *ẓarf* that arise from mother tongue interference and overgeneralization of target language rules. The third area concerned the theoretical framework underlying the diagnostic approach[19] adopted in this study.

[19] Badrun Kholid, Arif Rahman, and Lalu Ari Irawan, "Implementing Diagnostic Assessment in Designing Differentiated Learning for English Language Learning at the Junior High Schools," *JoLLS:*

Two theoretical foundations were identified as most relevant. The first is error analysis theory, which holds that learner errors are not random but systematic, arising from identifiable sources such as interlingual transfer, intralingual overgeneralization, and the influence of instructional context. This framework is particularly relevant to Arabic grammar learning among Indonesian learners, where errors in *fiʿl, adawāt al-istifhām*, and *ẓarf* may reflect predictable interference patterns from the learners' first language rather than simple gaps in knowledge, suggesting that a two-tier diagnostic test capable of surfacing the reasoning behind incorrect responses may offer more useful information than conventional correct-wrong scoring alone.

A complementary perspective is offered by conceptual change theory, which holds that learners develop alternative conceptions that tend to persist when new material conflicts with prior knowledge structures. This theory was originally developed in the context of science education,[20] so its application to language learning requires some adaptation. In the present study, it is used in a more limited sense, as a framework for understanding why certain students appear to retain incorrect understandings of Arabic grammar rules even after instruction, based on the assumption that grammatical rules can be viewed as rule-based knowledge structures comparable to scientific concepts. Arabic grammar rules such as the conjugation system of *fiʿl*, the semantic scope of *adawāt al-istifhām*, and the *iʿrāb* requirements of *ẓarf* may be resistant to revision among Indonesian learners when these rules conflict with the grammatical logic of their first language. This adapted use of the theory is acknowledged as a theoretical proposition that requires further empirical investigation.

At the Conceptual Framework Development stage, an initial design was formulated based on the results of the literature study. The item grid was developed by deriving learning indicators from the basic competencies of Arabic grammar at the 10th grade level, covering the materials of *fiʿl, adawāt al-istifhām*, and *ẓarf*. Based on the needs analysis, several anticipated misconception patterns were identified to guide item and distractor construction. In the *fiʿl* material, students were anticipated to confuse *fiʿl māḍī, muḍāriʿ*, and *amr* due to the absence of tense-based verb form changes in Indonesian.[21] In the *adawāt al-istifhām* material, students were anticipated to select inappropriate question words by

Journal of Language and Literature Studies 4, no. 2 (2024): 445–58, https://doi.org/https://doi.org/10.36312/jolls.v4i2.1934.

[20] Tsaniyah Nabilah Rachmawati and Z. A. Imam Supardi, "Analisis Model Conceptual Change Dengan Pendekatan Konflik Kognitif Untuk Mengurangi Miskonsepsi Fisika Dengan Metode Library Research," *PENDIPA Journal of Science Education* 5, no. 2 (2021): 133–42, https://doi.org/https://doi.org/10.33369/pendipa.5.2.133-142.

[21] Muneera Muftah, "Error Analysis in Second Language Acquisition (SLA): Types and Frequencies of Grammatical Errors of Simple Present and Past Tense in the Elicited Written Production Task of Arab EFL Undergraduate Learners," *Colombian Applied Linguistic Journal* 25, no. 1 (2023): 42–56, https://doi.org/https://doi.org/10.14483/22487085.19202.

overgeneralizing Indonesian question word functions onto Arabic. In the *ẓarf* material, students were anticipated to conflate *ẓarf zamān* and *ẓarf makān* due to unfamiliarity with *iʿrāb*-governed adverbial rules. These anticipated misconception patterns served as the basis for constructing distractors in both tiers of the instrument.

Prototyping Stage

The prototyping stage consisted of three prototypes in the formation of the test instrument. The development of Prototype 1 began with designing items based on the item grid that had been derived from the learning indicators and basic competencies. Ten two-tier diagnostic test items were developed in multiple-choice format, each with five answer options. The materials covered were *fiʿl* with a total of 4 items (items 1–4), *adawāt al-istifhām* with a total of 2 items (items 5–6), and *ẓarf* with a total of 4 items (items 7–10), complete with answer keys and assessment rubrics. The reason for the fewer items on *adawāt al-istifhām* compared to the other materials is that this material does not carry internal subdivisions, unlike *fiʿl* which has branches of *fiʿl māḍī, muḍāriʿ*, and *amr*, and *ẓarf* which has branches of *ẓarf zamān* and *ẓarf makān*. Each distractor in both tiers was constructed based on the anticipated misconception patterns identified in the preliminary research stage.

In Prototype 2, a self-evaluation was conducted on the developed instrument. The aspects reviewed included language clarity to avoid ambiguous phrasing, alignment between item indicators and learning indicators, completeness of item identity cards, appropriateness of answer options and reasoning options in each tier, and consistency of the answer key with the two-tier diagnostic framework. Revisions were made based on the results of this self-evaluation before proceeding with expert review.

In Prototype 3, expert review was conducted to assess the content and construct validity of the instrument. Content validation assessed three aspects, namely material suitability, language and writing quality, and relevance to the curriculum. Construct validation assessed three aspects, namely two-tier diagnostic test structure, distractor quality, and alignment with two-tier theory. The validators in this study were a lecturer in Arabic Language and Literature at UIN Maulana Malik Ibrahim Malang and a teacher at Madrasah Aliyah Assunniyyah Al-Jauhari Kencong.

The average construct validation score was 4.21 out of 5 (84.2%), meeting the Very Valid criterion. The scores across the three assessed aspects are presented in Table 2.

Table 2. Construct Validation Results

Assessment Aspects	Validator 1	Validator 2	Combined Average	Percentage	Criteria
Two-Tier Diagnostic Test Structure	4,75	4,00	4,38	87,5%	Very Valid
Distractor Quality	4,00	3,75	3,88	77,5%	Valid
Alignment with Two-Tier Theory	4,25	4,50	4,38	87,5%	Very Valid
Overall Average	**4,33**	**4,08**	**4,21**	**84,2%**	**Very Valid**

Meanwhile, the average content validation score was 4.42 out of 5 (88%), also meeting the Very Valid criterion, as presented in Table 3.

Table 3. Content Validation Results

Assessment Aspects	Validator 1	Validator 2	Combined average	Percentage	Criteria
Material Suitability	4,50	4,50	4,50	90%	Very Valid
Language and Writing Quality	4,75	4,25	4,50	90%	Very Valid
Relevance to the Curriculum	4,50	4,00	4,25	85%	Very Valid
Overall Average	**4,58**	**4,25**	**4,42**	**88%**	**Very Valid**

To assess the level of agreement between the two validators beyond percentage scores alone, inter-rater reliability was calculated using Gwet's AC1 coefficient. The results are presented in the following table.

Table 4. Inter-Rater Reliability Results

Validation Type	Pa	$\Sigma\pi_k(1-\pi_k)$	Pe	AC1	Category
Construct Validation	0,4167	0,5799	0,1450	0,3178	Fair
Content Validation	0,4167	0,5694	0,1424	0,3198	Fair

The AC1 value for construct validation was 0.3178 and for content validation was 0.3198, both falling within the *Fair* category (0.21–0.40) according to the criteria established by Landis and Koch (1977).[22] Although categorized as Fair, these values indicate that the validators' assessments were directionally consistent rather than arbitrary, and are considered acceptable for an instrument at this stage of development given that two experts assessing independently will naturally produce minor score variations.[23] When read alongside the validity percentages of 84.2% and 88%, these AC1 values suggest that the instrument meets

[22] J. Richard Landis and Gary G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics* 33, no. 1 (1977): 159–74, https://pubmed.ncbi.nlm.nih.gov/843571/.

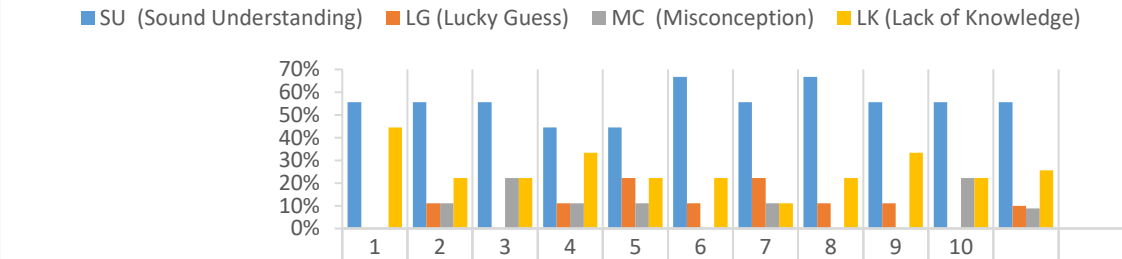
[23] K. L. Gwet, "The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters," in *Handbook of Inter-Rater Reliability* (STATAXIS Publishing Company, 2001), https://www.ncbi.nlm.nih.gov/books/NBK82266/.

threshold standards of both content quality and inter-rater consistency for this phase.

Several limitations should nonetheless be acknowledged. Inter-rater reliability in this study falls within the fair range, which represents a limitation, and the construct validity evidence reported here, based on expert judgment percentages rather than factor analytic or IRT approaches, should be regarded as initial rather than comprehensive evidence. The operationalization of construct validity in this study followed a content focused expert review model, in which validators assessed the alignment of item structure with two tier theory, distractor quality, and two-tier diagnostic test structure using a five-point Likert rubric. This approach is consistent with established practice in early-stage instrument development, while fuller construct validation, including factor analysis and criterion related validity, is identified as a necessary step for subsequent research.

Assessment Phase

The assessment phase was carried out through a small-scale trial and a large-scale trial, followed by item analysis and two-tier diagnostic test classification. The small-scale trial was conducted with 9 purposively selected students, composed of 3 high-ability, 3 medium-ability, and 3 low-ability students. The item validity analysis showed that all 10 items were declared valid, with correlation coefficients ranging from 0.64 to 0.93, all exceeding the r-table value of 0.5822 at a significance level of 5% with n=9. Reliability showed Cronbach's Alpha of 0.93, indicating high internal consistency. Seven items (70%) demonstrated good discriminating power and two items (20%) demonstrated very good discriminating power. In terms of difficulty level, seven items were classified as moderate and three items were classified as easy. The two-tier diagnostic test results from the small-scale trial showed that 56% of students were classified as SU, 10% as LG, 9% as MC, and 26% as LK. These preliminary results indicated sufficient item quality to proceed to the large-scale trial.



	1	2	3	4	5	6	7	8	9	10	Total
SU (Sound Understanding)	56%	56%	56%	44%	44%	67%	56%	67%	56%	56%	56%
LG (Lucky Guess)	0%	11%	0%	11%	22%	11%	22%	11%	11%	0%	10%
MC (Misconception)	0%	11%	22%	11%	11%	0%	11%	0%	0%	22%	9%
LK (Lack of Knowledge)	44%	22%	22%	33%	22%	22%	11%	22%	33%	22%	26%

Figure 1. Distribution of Student Response Categories in the Small-Scale Trial

After the small-scale trial showed positive results, a large-scale trial was conducted with the full class population of 31 students. The item validity analysis showed that all 10 items were declared valid, with correlation coefficients ranging from 0.41 to 0.83, all exceeding the r-table value of 0.3009 at a significance level of 5% with n=31. The results of the item validity analysis are presented in Table 5.

Table 5. Item Validity Results in the Large-Scale Trial

Item	R-calculated	R-table	Criteria
1	0.59	0.3009	Valid
2	0.80	0.3009	Valid
3	0.69	0.3009	Valid
4	0.52	0.3009	Valid
5	0.66	0.3009	Valid
6	0.66	0.3009	Valid
7	0.41	0.3009	Valid
8	0.77	0.3009	Valid
9	0.83	0.3009	Valid
10	0.70	0.3009	Valid

The reliability analysis using Cronbach's Alpha[24] produced a coefficient of 0.86, indicating adequate internal consistency. The reliability calculation details are presented below:

Table 6. Reliability Analysis Results in the Large-Scale Trial

Question Number	**Variants**
1	0,72
2	0,66
3	0,85
4	0,85
5	0,69
6	0,83
7	0,57
8	0,94
9	0,81
10	0,80
Number of Item Variants	**7,71**
Total Score Variant	**34,1**
Reliability	**0,86**

The difficulty level analysis showed that 80% of items (8 out of 10) were classified as moderate (P = 0.30–0.70), which is considered appropriate for a two-tier diagnostic test. Two items were classified as easy (P > 0.70). No items were classified as difficult, indicating a reasonable distribution for measuring student abilities at various levels. The difficulty level results are presented below:

[24] Muhammad Amirrudin, Khoirunnisa Nasution, and Supahar Supahar, "Effect of Variability on Cronbach Alpha Reliability in Research Practice," *Jurnal Matematika, Statistika Dan Komputasi* 17, no. 2 (December 23, 2020): 223–30, https://doi.org/10.20956/jmsk.v17i2.11655.

Table 7. Difficulty Level Analysis Results in the Large-Scale Trial

Item	Difficulty Index (P)	Category
1	0,74	Easy
2	0,73	Easy
3	0,44	Moderate
4	0,39	Moderate
5	0,60	Moderate
6	0,66	Moderate
7	0,52	Moderate
8	0,58	Moderate
9	0,58	Moderate
10	0,53	Moderate

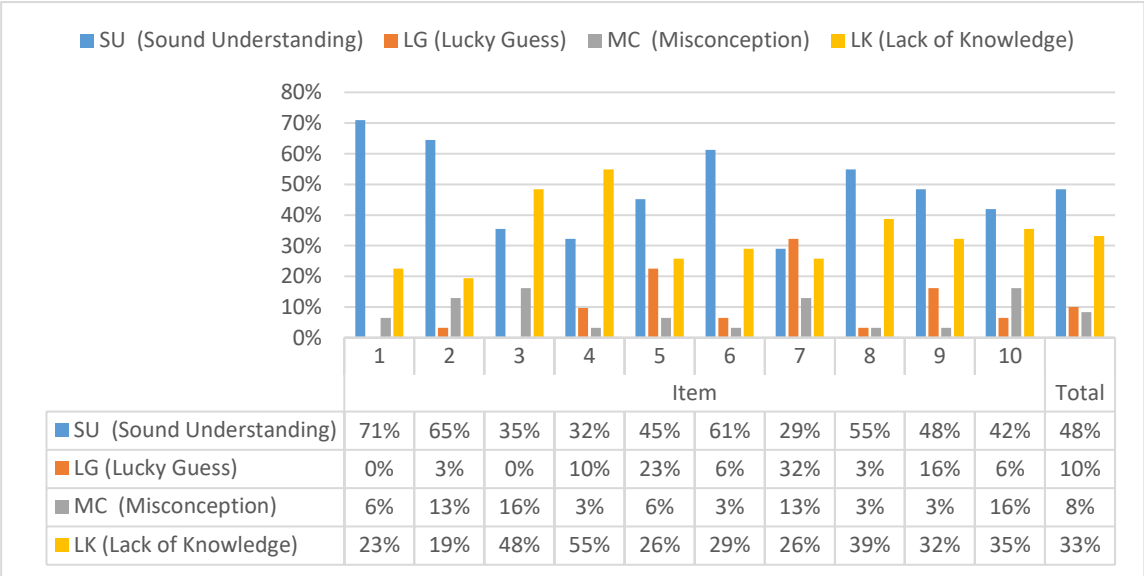
The discriminating power analysis showed that 80% of items demonstrated good discriminating power ($0.40 < DP \leq 0.70$), with item 9 showing very good discriminating power (DP = 0.80). One item, item 7, recorded low discriminating power (DP = 0.16) and is identified as requiring revision in subsequent development phases. The following table shows the discriminating power results:

Table 8. Discriminating Power Analysis Results in the Large-Scale Trial

Item	Upper-Class Average	Lower-Class Average	DP	Category
1	1.88	1.00	0.44	Good
2	2.00	0.79	0.61	Good
3	1.41	0.21	0.60	Good
4	1.18	0.29	0.45	Good
5	1.65	0.64	0.50	Good
6	1.76	0.79	0.49	Good
7	1.18	0.86	0.16	Poor
8	1.65	0.57	0.54	Good
9	1.88	0.29	0.80	Very Good
10	1.59	0.43	0.58	Good

Based on validity, reliability, difficulty level, and discriminating power analyses, the two-tier diagnostic test instrument showed preliminary psychometric adequacy for the purposes of this study. All items were declared valid at the item level, the instrument showed acceptable internal consistency (Cronbach's alpha = 0.86), and most items demonstrated satisfactory difficulty and discriminating power distributions, although these findings should be interpreted as preliminary indicators rather than evidence of established psychometric stability, given that the analyses were conducted with a small, single school sample. Item 7 is flagged as a candidate for revision in subsequent development phases due to its poor discriminating power (DP = 0.16), which suggests it may not adequately differentiate between higher and lower performing students. More robust psychometric evidence, including Exploratory Factor Analysis, Confirmatory Factor Analysis, Rasch modeling, or multi-site trials with larger and more diverse samples, was not feasible at this stage and represents a direction for future validation work.

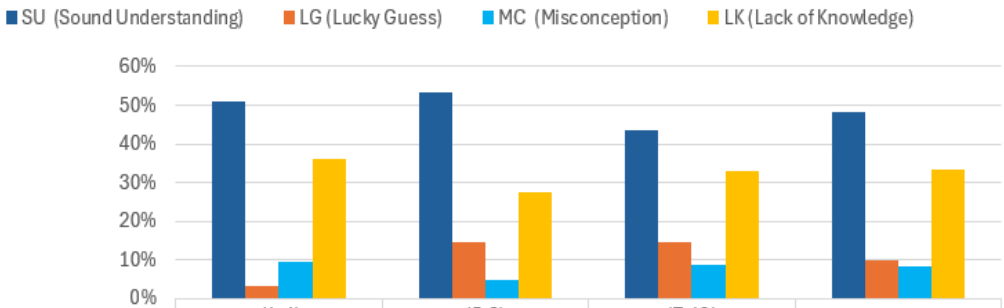
The two-tier diagnostic test analysis was conducted to classify student responses across the four diagnostic categories. The overall results showed that 48% of students were classified as SU, 10% as LG, 8% as MC, and 33% as LK. More detailed information is presented in the figures below.



	1	2	3	4	5	6	7	8	9	10	Total
	Item										
SU (Sound Understanding)	71%	65%	35%	32%	45%	61%	29%	55%	48%	42%	48%
LG (Lucky Guess)	0%	3%	0%	10%	23%	6%	32%	3%	16%	6%	10%
MC (Misconception)	6%	13%	16%	3%	6%	3%	13%	3%	3%	16%	8%
LK (Lack of Knowledge)	23%	19%	48%	55%	26%	29%	26%	39%	32%	35%	33%

Figures 2. Distribution of Student Response Categories in the Large -Scale Trial

In this study, the test covered three materials. The first material was *fiʿl*, which included items 1 to 4 and covered *fiʿl māḍī, muḍāriʿ*, and *amr*. The second material was *adawāt al-istifhām*, which included items 5 to 6. The third material was *ẓarf*, which included items 7 to 10 and covered *ẓarf zamān* and *ẓarf makān*. The smaller number of items for *adawāt al-istifhām* reflects the absence of internal subdivisions in this material compared to the other two. The distribution of student response categories across the three materials is presented in the following diagram.



	(1-4) FI'L	(5-6) ADAWĀT AL-ISTIFHĀM	(7-10) ZARF	Total
SU (Sound Understanding)	51%	53%	44%	48%
LG (Lucky Guess)	3%	15%	15%	10%
MC (Misconception)	10%	5%	9%	8%
LK (Lack of Knowledge)	36%	27%	33%	33%

Figure 3. Distribution of Student Response Categories by Material in the Large -Scale Trial

The most notable finding is that 33% of students overall fell into the LK category, indicating an absence of both conceptual understanding and consistent reasoning. The highest LK percentage was found in *fiʿl* (36%), followed by *ẓarf* (33%) and *adawāt al-istifhām* (27%), which is consistent with the anticipated difficulty patterns identified in the preliminary research stage, where *ẓarf* was expected to be the most challenging material due to the absence of *iʿrāb*-equivalent rules in Indonesian. The MC category, while relatively small in overall percentage (8%), is considered the most pedagogically significant finding, as students in this category hold structured yet erroneous beliefs that may require targeted conceptual restructuring rather than mere repetition or foundational re-teaching.

The findings of this study indicate that misconceptions in Arabic grammar learning are not distributed randomly across materials but tend to cluster around topics where Arabic and Indonesian grammatical structures diverge most significantly, particularly *fiʿl, adawāt al-istifhām*, and *ẓarf*. This pattern is consistent with error analysis theory as proposed by Corder,[25] which holds that learners' errors are systematic rather than arbitrary, and are rooted in prior linguistic knowledge, mother tongue interference,[26] and rule overgeneralization.[27]

However, these structural differences represent only one plausible explanation among several for the error patterns identified. Other factors, including instructional quality, limited exposure to authentic Arabic text, memorization-based learning approaches, and the way grammar rules are presented in available textbooks,[28] may also contribute to learner difficulties. As the diagnostic categories produced by this instrument describe patterns of understanding rather than causal mechanisms, the interpretation of findings here distinguishes between observed diagnostic classifications and the hypothesized sources underlying them.[29]

A complementary theoretical perspective is offered by conceptual change theory, which holds that learners develop alternative conceptions that persist when new material conflicts with prior knowledge structures. It should be noted

[25] Ana Sari Putri, "Analisis Kesalahan Berbahasa Stephen Pit Corder Dalam Praktek Percakapan Berbahasa Arab Siswi Kelas IX SMPIT Baitussalam Prambanan Tahun Ajaran 2015/2016" (UIN Sunan Kalijaga Yogyakarta, 2016), https://digilib.uin-suka.ac.id/id/eprint/20898/.

[26] Zohra Benamor and Nawel Benmostefa, "Diagnostic Testing as an Effective Tool for Learning Case Study : A Study of Difficulties Encountered by EFL Learners When Using Tenses in English," *Aleph: Langues et Cultures Linguistiques à l'oeuvre* 9, no. 3 (2025): 1–22, https://aleph.edinum.org/6428.

[27] Salem Khalil Al Aqtash and Omran Ahmad Musleh, "The Effect of Concept Mapping on Arabic Grammar Proficiency : Al Ain University Students in the United Arab Emirates," *Universal Journal of Educational Research* 8, no. 9 (2020): 4089–96, https://doi.org/10.13189/ujer.2020.080934.

[28] Sri Wahyuni Mohamad and Muhammad Nur Iman, "Analisis Kelayakan Penyajian Pada Buku Teks Bahasa Arab Kelas X Madrasah Aliyah KMA 183 Tahun 2019," *Al-Muyassar: Journal of Arabic Education* 2, no. 1 (2023): 19–32, https://doi.org/https://doi.org/10.31000/al-muyassar.v2i1.6573.

[29] Jiuliang Li and Noriko Iwashita, "Examining The Effects of A Diagnostic Language Test on Learning," *The Journal of Educational Research* 118, no. 2 (2025): 156–68, https://doi.org/https://doi.org/10.1080/00220671.2025.2454682.

that this theory was originally developed in the context of science education, and its direct application to language learning requires some adaptation.[30] In the present study, it is used in a more limited sense, namely as a framework for understanding why certain students retain incorrect understandings of Arabic grammar rules even after instruction. The two-tier diagnostic test developed in this study does not investigate the origins of those misunderstandings but is designed to identify and map them across four categories, namely Sound Understanding (SU), Lucky Guess (LG), Misconception (MC), and Lack of Knowledge (LK), so that educators may subsequently address them through more targeted instructional approaches.

This four-category classification provides more diagnostic specificity than a simple correct–incorrect scoring system. Rather than merely indicating whether a student answered correctly, the classification reveals the nature of each student's understanding. Students classified as LG selected the correct answer in the first tier but were unable to provide adequate justification in the second tier, suggesting that their correct answers may reflect familiarity or guessing rather than conceptual grasp. Students in the MC category selected an incorrect first-tier answer but provided correct reasoning in the second tier, indicating that some conceptual understanding has begun to develop but has not yet been accurately applied to formal identification. Students in the LK category were unable to answer either tier adequately, suggesting an absence of initial comprehension rather than the presence of a misconception in the strict sense.[31]

In order to further validate these classifications, semi-structured interviews were conducted with nine students after the test administration. Three students were selected purposively from each of the three diagnostic categories (LG, MC, and LK), with selection based on the three highest scores within each category. Each session focused on eliciting students' reasoning behind their answers, with particular attention to whether verbal explanations were consistent with the diagnostic category assigned by the instrument. The interview data were then used to examine whether the response patterns and diagnostic categories produced by the two-tier instrument were consistent with the verbal explanations provided by the students. It is acknowledged that the small number of interview participants (n=9) limits the generalizability of these qualitative findings. They are presented here as illustrative triangulation evidence rather than as independent validation of the instrument's classifications.

[30] Anita Dewi Utami, M Zainudin, and Leilita Anggraini, "Perubahan Konseptual Siswa Dalam Memahami Konsep Fungsi Ditinjau Dari Gaya Kognitif Field Dependent Dan Field Independent Dalam Pembelajaran Daring," *Educatif : Journal of Education Research* 2, no. 4 (2020): 1–16.

[31] Anam and Gumilar, "A Systematic Literature Review of Tier Tests as Diagnostic Tools in Specific Areas of Science."

In the LG category, the three participants showed a consistent pattern of selecting correct first-tier answers without being able to explain the grammatical basis for their choices. In the *fiʿl māḍī* items, Participant 1 reported choosing كَتَبَتْ because the word sounded familiar, without knowledge that it is a feminine *fiʿl māḍī* marked by the *tāʾ al-taʾnīṡ al-sākinah*. Participant 3 similarly admitted to guessing ذَهَبَ without any awareness of the formal features of *fiʿl māḍī*. In the *adawāt al-istifhām* items, Participant 2 recognized مَتَى as an interrogative word but could not specify that its function is to ask about time. On the ẓarf items, students in this group generally could not distinguish *ẓarf makān* from *ẓarf zamān* conceptually, despite showing partial vocabulary familiarity. These patterns are consistent with the LG classification, in that correct surface-level responses were not accompanied by adequate conceptual understanding.

In the MC category, the three participants showed the opposite pattern: incorrect first-tier answers alongside correct second-tier reasoning. Participant 5 reported retaining only the definition of *fiʿl māḍī* and *fiʿl amr* from classroom instruction, without having retained the formal markers of each. Participant 6 acknowledged that *fiʿl māḍī* forms tend to be confused with one another, despite understanding that *fiʿl māḍī* refers to past events. On the ẓarf items, Participant 4 understood that *ẓarf makān* relates to place and *ẓarf zamān* to time but could not recall which vocabulary items belong to each category. Participant 5 selected الصبح as a *ẓarf zamān* based solely on its meaning, without considering that classification also depends on syntactic position and *iʿrāb* as a *manṣūb mafʿūl fīh*. Participant 6 admitted to not having memorized the specific functions of individual interrogative words, resulting in an incorrect first-tier answer despite partially correct second-tier justification.

In the LK category, the three participants demonstrated the most pervasive absence of understanding, with LK responses comprising 70–80% of their answer profiles. Participants 7 and 8 reported having no knowledge of the formal features of *fiʿl amr* or *fiʿl māḍī* and acknowledged guessing across all items. Both participants were also unfamiliar with the forms and specific functions of the *adawāt al-istifhām*. Participant 9 acknowledged that the ẓarf material had been covered in class but stated that insufficient attention during instruction had resulted in no retained understanding. These findings suggest that the difficulties in this group reflect an absence of initial comprehension rather than misconceptions in the strict sense, consistent with their LK classification.

The diagnostic categories also carry potential implications for instructional practice. For students in the LK category, instruction may benefit from building foundational understanding through explicit explanation and, where applicable,

contrastive analysis between Arabic and Indonesian grammatical structures,[32] given that an established conceptual framework does not yet appear to have formed. For students in the MC category, activities that surface and examine already formed understandings may be more effective than re explanation alone. Since some conceptual base appears to already be present but has not been accurately applied, the pedagogical need may be to help students restructure existing understanding rather than simply acquire new knowledge. For students in the LG category, structured practice that reinforces the connection between correct answers and their grammatical rationale may help consolidate understanding that currently appears incomplete. These instructional implications are proposed as tentative directions informed by the diagnostic classification, and their effectiveness would need to be evaluated through further classroom-based research.[33]

CONCLUSION

This study produced a ten-item two-tier diagnostic test instrument covering the Arabic grammar materials of *fiʿl, adawāt al-istifhām*, and *ẓarf*, developed through the Plomp model. Conducted at MA Ma'arif NU Kencong, the instrument received Very Valid ratings from both content validation (88%) and construct validation (84.2%). Item analysis in the large-scale trial showed that all 10 items were valid, with correlation coefficients ranging from 0.41 to 0.83, a Cronbach's Alpha of 0.86, 80% of items at moderate difficulty level, and 80% of items demonstrating good discriminating power.

The two-tier diagnostic test classification showed that 48% of students were classified as Sound Understanding (SU), 10% as Lucky Guess (LG), 8% as Misconception (MC), and 33% as Lack of Knowledge (LK). The relatively high proportion of LK across all three materials, particularly in *fiʿl* (36%) and *ẓarf* (33%), suggests that a substantial number of students may benefit from foundational re-teaching rather than misconception-targeted intervention alone. The MC category, while smaller in percentage, identified students whose errors appear to reflect structured yet inaccurate understandings, which may require different instructional approaches than those appropriate for LK students.

Several limitations should be noted. The study was conducted at a single school with a limited sample of 9 students in the small-scale trial and 31 students in the large-scale trial, which is insufficient for establishing robust psychometric properties. Inter rater reliability values, while considered acceptable for this stage

[32] Ahmad Maghfurin et al., "Implementation of Behavioral and Cognitive Theories in The Arabic Language Learning" *Al Mi'yar: Jurnal Ilmiah Pembelajaran Bahasa Arab Dan Kebahasaaraban* 6, no. 2 (2023): 325–42, https://doi.org/10.35931/am.v6i2.2609.

[33] Hossein Farhady and Merve Selcuk, "Classroom-Based Diagnostic Assessment Practices of EFL Instructors," *Iranian Journal of Language Teaching Research* 10, no. 2 (2022): 77–94, https://doi.org/10.30466/ijltr.2022.121184.

of development, remained in the fair range, and Item 7 recorded poor discriminating power (DP = 0.16), indicating a need for revision. More comprehensive psychometric analyses, including Exploratory Factor Analysis, Confirmatory Factor Analysis, Rasch modeling, criterion related validity, and test-retest reliability, were not conducted and represent gaps to be addressed in future work. The qualitative interview data, drawn from nine students only, should be regarded as illustrative rather than confirmatory. Further research involving larger and more diverse samples from multiple schools, as well as longitudinal testing to examine instrument stability, is therefore recommended.

References

Alhafni, Bashar, Go Inoue, Christian Khairallah, and Nizar Habash. "Advancements in Arabic Grammatical Error Detection and Correction : An Empirical Investigation." In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6430–48. Association for Computational Linguistics, 2023. https://doi.org/https://doi.org/10.18653/v1/2023.emnlp-main.396.

Alothman, Ameerah, and Abdulmalik Alsalman. "An Arabic Grammar Auditor Based on Dependency Grammar." *Hindawi: Advances in Human-Computer Interaction* 2020, no. 1 (2020): 1–10. https://doi.org/https://doi.org/10.1155/2020/8856843.

Amirrudin, Muhammad, Khoirunnisa Nasution, and Supahar Supahar. "Effect of Variability on Cronbach Alpha Reliability in Research Practice." *Jurnal Matematika, Statistika Dan Komputasi* 17, no. 2 (December 23, 2020): 223–30. https://doi.org/10.20956/jmsk.v17i2.11655.

Anam, Rifat Shafwatul, and Surya Gumilar. "A Systematic Literature Review of Tier Tests as Diagnostic Tools in Specific Areas of Science." *Tadris: Jurnal Keguruan Dan Ilmu Tarbiyah* 9, no. 1 (June 29, 2024): 255–70. https://doi.org/10.24042/tadris.v9i1.17717.

Aqtash, Salem Khalil Al, and Omran Ahmad Musleh. "The Effect of Concept Mapping on Arabic Grammar Proficiency : Al Ain University Students in the United Arab Emirates." *Universal Journal of Educational Research* 8, no. 9 (2020): 4089–96. https://doi.org/10.13189/ujer.2020.080934.

Azizah, Nur Laeli, Lina Mahardiani, and Sri Yamtinah. "Analisis Miskonsepsi Dengan Tes Diagnostik Two-Tier Multiple Choice Dan In-Depth Interview Pada Materi Asam Basa." *Jurnal Pendidikan Kimia* 11, no. 2 (2022): 168–77. https://doi.org/10.20961/jpkim.v11i2.60345.

Benamor, Zohra, and Nawel Benmostefa. "Diagnostic Testing as an Effective Tool for Learning Case Study : A Study of Difficulties Encountered by EFL Learners When Using Tenses in English." *Aleph: Langues et Cultures Linguistiques à l'oeuvre* 9, no. 3 (2025): 1–22. https://aleph.edinum.org/6428.

Daud, Wan Ab Aziz Wan, Mohammad Taufiq Abdul Ghani, Ahmad Abdul Rahman, Akashah Mohd, Yusof Mohamad, Hadhrami Ab Ghani, and Zahazan Mohamed. "Exploring the Challenges and Strategies of Learning Arabic Language among Primary School Teachers." *Forum for Linguistic Studies* 7, no. 2 (2025): 94–103. https://doi.org/10.30564/fls.v7i2.8002.

Farhady, Hossein, and Merve Selcuk. "Classroom-Based Diagnostic Assessment Practices of EFL Instructors." *Iranian Journal of Language Teaching Research* 10, no. 2 (2022): 77–94. https://doi.org/10.30466/ijltr.2022.121184.

Gurel, Derya Kaltakci, Ali Eryılmaz, and Lillian Christie Mcdermott. "A Review and Comparison of Diagnostic Instruments to Identify Students ' Misconceptions in Science." *Eurasia Journal of Mathematics, Science & Technology Education* 11, no. 5 (2015): 989–1008. https://doi.org/10.12973/eurasia.2015.1369a.

Gwet, K. L. "The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters." In *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company, 2001. https://www.ncbi.nlm.nih.gov/books/NBK82266/.

Handjoko Permana, Yetty Supriyati, Dewi Muliyati, Ananda Ayu Dewi Sekartaji, Marie Paz E. Morales, and Diah Ambarwulan. "Development of a Two-Tier Diagnostic Test for The Analysis of Momentum and Impulse Learning Difficulties in Students." *Jurnal Penelitian & Pengembangan Pendidikan Fisika* 9, no. 2 (December 31, 2023): 237–44. https://doi.org/10.21009/1.09206.

Kholid, Badrun, Arif Rahman, and Lalu Ari Irawan. "Implementing Diagnostic Assessment in Designing Differentiated Learning for English Language Learning at the Junior High Schools." *JoLLS: Journal of Language and Literature Studies* 4, no. 2 (2024): 445–58. https://doi.org/https://doi.org/10.36312/jolls.v4i2.1934.

Kismiati, Dyah Aniza, and Leonard R Hutasoit. "Teacher Misconceptions : A Phenomenon of the Lack of Knowledge in Science Subjects." *JPPIPA: Jurnal Penelitian Pendidikan IPA* 10, no. 14 (2024): 3493–3500. https://doi.org/10.29303/jppipa.v10i6.7226.

Landis, J. Richard, and Gary G. Koch. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33, no. 1 (1977): 159–74. https://pubmed.ncbi.nlm.nih.gov/843571/.

Levy-feldman, Irit. "The Role of Assessment in Improving Education and Promoting Educational Equity." *Education Sciences* 15, no. 2 (2025): 1–11. https://doi.org/https://doi.org/10.3390/educsci15020224.

Li, Jiuliang, and Noriko Iwashita. "Examining The Effects of A Diagnostic Language Test on Learning." *The Journal of Educational Research* 118, no. 2 (2025): 156–68. https://doi.org/https://doi.org/10.1080/00220671.2025.2454682.

Maghfurin, Ahmad, Irfatin Maisaroh, M Aziz Himawan Akbar, and M Rizki. "Implementation of Behavioral and Cognitive Theories in The Arabic Language Learning." *Al Mi'yar: Jurnal Ilmiah Pembelajaran Bahasa Arab Dan Kebahasaaraban* 6, no. 2 (2023): 325–42. https://doi.org/10.35931/am.v6i2.2609.

Mamnunah, Mamnunah, Maman Abdurrahman, and Asep Sopian. "The Error Analysis of Arabic Grammar in The Kalamuna Book." *Arabi: Journal of Arabic Studies* 6, no. 2 (2021): 158–66. https://doi.org/http://dx.doi.org/10.24865/ajas.v6i2.351.

Maulidiansyah, Deka, Inarsih Meutia, and Edi Istiyono. "Computer-Based Two-Tier Diagnostic Test to Identify Critical Thinking Skills in Optical Instrument." In *Proceedings of the 6th International Seminar on Science Education (ISSE 2020)*, 541:413–18, 2021. https://doi.org/10.2991/assehr.k.210326.059.

Mohamad, Sri Wahyuni, and Muhammad Nur Iman. "Analisis Kelayakan Penyajian Pada Buku Teks Bahasa Arab Kelas X Madrasah Aliyah KMA 183 Tahun 2019." *Al-*

Muyassar: Journal of Arabic Education 2, no. 1 (2023): 19–32. https://doi.org/https://doi.org/10.31000/al-muyassar.v2i1.6573.

Muftah, Muneera. "Error Analysis in Second Language Acquisition (SLA): Types and Frequencies of Grammatical Errors of Simple Present and Past Tense in the Elicited Written Production Task of Arab EFL Undergraduate Learners." *Colombian Applied Linguistic Journal* 25, no. 1 (2023): 42–56. https://doi.org/https://doi.org/10.14483/22487085.19202.

Omer, Neimat Idris Moh. Saeed. "Error Analysis of Grammatical Features in Saudi Undergraduate EFL Writing." *ESIC* 8, no. 1 (2024): 483–98.

Putri, Ana Sari. "Analisis Kesalahan Berbahasa Stephen Pit Corder Dalam Praktek Percakapan Berbahasa Arab Siswi Kelas IX SMPIT Baitussalam Prambanan Tahun Ajaran 2015/2016." UIN Sunan Kalijaga Yogyakarta, 2016. https://digilib.uin-suka.ac.id/id/eprint/20898/.

Rachmawati, Tsaniyah Nabilah, and Z. A. Imam Supardi. "Analisis Model Conceptual Change Dengan Pendekatan Konflik Kognitif Untuk Mengurangi Miskonsepsi Fisika Dengan Metode Library Research." *PENDIPA Journal of Science Education* 5, no. 2 (2021): 133–42. https://doi.org/https://doi.org/10.33369/pendipa.5.2.133-142.

Rahmawati, Kartika Eva. "Using Iconicity in Two-Tier Grammar Test for Adult Learners: A Cognitive Linguistic Approach." *ELECTRUM (The English Language and Education Spectrum)* 2, no. 2 (July 2022): 1–7.

Ramadhan, Gilang, Pratiwi Dwijananti, and Siti Wahyuni. "Analisis Kemampuan Berpikir Tingkat Tinggi (High Order Thinking Skills) Menggunakan Instrumen Two Tier Multiple Choice Materi Konsep Dan Fenomena Kuantum Siswa SMA Di Kabupaten Cilacap." *UPEJ: Unnes Physics Education Journal* 7, no. 3 (2018): 85–90. http://journal.unnes.ac.id/sju/index.php/upej.

Ridho, Muhammad Rifqi, and Abu Bakar. "Signifikansi Gramatika Arab Dalam Menginterpretasikan Al-Qur'an: Tinjauan Ilmu Nahwu, Ilmu Sharaf, Dan Ilmu Balaghah." *AL-MAZAYA Jurnal Pendidikan Bahasa Arab* 3, no. 1 (2025): 30–44.

Romadhon, Ilham Fatkhu. "Analisis Kesalahan Berbahasa Arab Pada Ketrampilan Berbicara Dalam Debat Bahasa Arab Di Indonesia." *Indonesian Journal of Arabic Education and Learning* 1, no. 1 (2025): 35–55. https://doi.org/10.58363/ijael.

Treagust, David F. "Diagnostic Assessment in Science as A Means to Improving Teaching, Learning and Retention." In *UniServe Science Assessment Symposium Proceedings*, 1–9. Proceedings of The Australian Conference on Science and Mathematics Education, 2012.

Tukiyo, Mukhtar Efendi, Everhard Markiano Solissa, Ika Yuniwati, and Syatria Adymas Pranajaya. "The Development of A Two-Tier Diagnostic Test to Detect Student's Misconceptions in Learning Process." *MUDIR (Jurnal Manajemen Pendidikan)* 5, no. 1 (2023): 92–96. https://doi.org/10.55352/mudir.

Utami, Anita Dewi, M Zainudin, and Leilita Anggraini. "Perubahan Konseptual Siswa Dalam Memahami Konsep Fungsi Ditinjau Dari Gaya Kognitif Field Dependent Dan Field Independent Dalam Pembelajaran Daring." *Educatif : Journal of Education Research* 2, no. 4 (2020): 1–16.