Asalibuna
مجلة اللغة العربية وتعليمها

E-ISSN: 2656-3533
P-ISSN: 2615-5885
Volume 10, Issue 01, 2026
DOI:

Redefining Arabic Proficiency Testing: A Flexible and Standardized Online TOAFL System

Akhmad Jazuli Afandi; Gigih Wahyu Pratomo.
UIN Syekh Wasil Kediri, Indonesia

Abstract:

This study aimed to develop an online Test of Arabic as a Foreign Language (TOAFL) platform based on a Learning Management System (LMS) and to construct a comprehensive item bank at UIN Syekh Wasil Kediri. TOAFL functions as a standardized instrument for assessing Arabic language proficiency among non-native speakers and enables universities to compare students' proficiency across diverse academic backgrounds. The institutional need arose from a substantial infrastructure imbalance. The Language Center had only one laboratory with 20 computers, whereas each testing period served approximately 1,000 to 1,500 examinees. This condition reduced the efficiency and accessibility of test administration. The study employed a Research and Development (R&D) design using the 4-D model, which comprises Define, Design, Develop, and Disseminate. Subject-matter and media experts validated the product to assess its feasibility and functionality. The study produced an LMS-based online TOAFL platform that supports a Test from Anywhere (TfA) system and integrates a structured item bank. The platform enables examinees to complete the test remotely without attending the UIN Syekh Wasil Kediri Language Center in person, thereby expanding access, reducing dependence on institutional computer laboratories, and improving the efficiency of TOAFL administration.

Keywords: Test of Arabic as a Foreign Language; Learning Management System; Research and Development; Test from Anywhere; Arabic Language Proficiency.

Submitted : December 20th 2025
Revised : July 31st 2026
Accepted : August 14th 2026

Correspondence Author:
Akhmad Jazuli Afandi
jazzull212@gmail.com

CC BY ND

Open Access Journal
Availabel at: https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/asalibuna

Asalibuna
مجلة اللغة العربية وتعليمها

E-ISSN: 2656-3533
P-ISSN: 2615-5885
Volume 10, Issue 02, 2026
DOI:

Mendefinisikan Ulang Ujian Kemahiran Bahasa Arab: Sistem TOAFL Daring yang Fleksibel dan Terstandarisasi

Akhmad Jazuli Afandi; Gigih Wahyu Pratomo.
UIN Syekh Wasil Kediri, Indonesia

Abstrak:

Penelitian ini bertujuan untuk mengembangkan platform ujian daring *Test of Arabic as a Foreign Language* (TOAFL) berbasis *Learning Management System* (LMS) dan menyusun bank soal yang komprehensif di UIN Syekh Wasil Kediri. TOAFL berfungsi sebagai instrumen terstandarisasi untuk menilai kemahiran bahasa Arab bagi penutur non-jati dan memungkinkan universitas untuk membandingkan kemahiran mahasiswa dari berbagai latar belakang akademik. Kebutuhan institusional ini muncul akibat ketidakseimbangan infrastruktur yang substansial. Pusat Bahasa hanya memiliki satu laboratorium dengan 20 komputer, padahal setiap periode ujian melayani sekitar 1.000 hingga 1.500 peserta. Kondisi ini mengurangi efisiensi dan aksesibilitas pelaksanaan ujian. Penelitian ini menggunakan desain *Research and Development* (R&D) dengan model 4-D, yang terdiri dari *Define* (Pendefinisian), *Design* (Perancangan), *Develop* (Pengembangan), dan *Disseminate* (Penyebarluasan). Pakar materi dan media memvalidasi produk ini untuk menilai kelayakan dan fungsionalitasnya. Penelitian ini menghasilkan platform TOAFL daring berbasis LMS yang mendukung sistem *Test from Anywhere* (TfA) dan mengintegrasikan bank soal yang terstruktur. Platform ini memungkinkan peserta untuk menyelesaikan ujian dari jarak jauh tanpa harus hadir langsung di Pusat Bahasa UIN Syekh Wasil Kediri, sehingga memperluas akses, mengurangi ketergantungan pada laboratorium komputer institusi, dan meningkatkan efisiensi pelaksanaan TOAFL.

Kata Kunci: *Test of Arabic as a Foreign Language*, *Learning Management System*, *Research and Development*, *Test from Anywhere*, Kemahiran Bahasa Arab.

Submitted : December 20th 2025
Revised : July 31st 2026
Accepted : August 14th 2026

Correspondence Author:
Akhmad Jazuli Afandi
jazzull212@gmail.com

CC BY ND

Open Access Journal
Availabel at: https://jurnalfaktarbiyah.iainkediri.ac.id/index.php/asalibuna

INTRODUCTION

Universities seeking to strengthen their global presence in an increasingly interconnected world must prioritize students' language proficiency. For institutions that integrate Arabic into their curricula or maintain academic engagement with Arabic-speaking regions, the Test of Arabic as a Foreign Language (TOAFL) constitutes an important assessment instrument. TOAFL provides a standardized indicator of Arabic proficiency and helps determine whether students possess the linguistic competence required to participate effectively in academic and professional environments in which Arabic plays a significant role (Nurdianto et al. 2025).

TOAFL constitutes a standardized instrument for assessing the Arabic language proficiency of non-native speakers (Alhawary 2019). Despite the importance of such assessment, institutions still lack a standardized online TOAFL platform, which limits the equitable and optimal measurement of Arabic proficiency. This gap establishes the need to develop a standardized assessment system that supports students' Arabic proficiency and advances the university's objective of expanding global partnerships.

In line with the internationalization of higher education, standardized Arabic proficiency assessment through TOAFL serves as a strategic instrument for strengthening students' academic competitiveness. TOAFL certification verifies the Arabic competence required for international research collaboration, student exchange programs, and academic and professional activities in Arabic-speaking contexts (Qodri 2020). Implementing TOAFL also reflects an institution's commitment to global educational quality standards through the use of a widely recognized language proficiency assessment instrument (Afyuddin et al. 2023). In addition to strengthening academic quality and institutional reputation, TOAFL prepares graduates for international careers in diplomacy, business, translation, and international organizations that require verified Arabic competence (Ulya et al. 2022).

This orientation aligns with the vision of UIN Syekh Wasil Kediri to become an internationally distinguished university in the development of scholarship, Islamic studies, and Indonesian values (Hamid and Charis 2021). In its implementation, TOAFL at UIN Kediri functions as a language proficiency test (*ikhtibār al-kafā'ah fī al-lughah al-'Arabiyyah*) that remains independent of any particular curriculum, textbook, or instructional program because it measures general Arabic competence rather than specific learning outcomes (Hasan 2018; Al Muchith and Nely Rahmawati Zaimah 2023).

This distinction positions TOAFL as a proficiency test rather than a curriculum-based achievement test. A team of lecturers with expertise in Arabic language education developed the TOAFL instrument at UIN Syekh Wasil Kediri. The instrument consists of 140 multiple-choice items covering three principal components, namely *istimā'* (50 items), *qawā'id* (40 items), and *qirā'ah* (50 items), together with five *insyā'* items.

The test content covers grammar (*nahwu* and *ṣarf*), vocabulary, terminology, *tarākīb*, and other linguistic dimensions to comprehensively assess Arabic competence. The *istimā'* section measures the ability to understand spoken discourse, the *qawā'id* section assesses mastery of Arabic grammatical structures, and the qirā'ah section focuses on reading comprehension (*al-qirā'ah al-ṣāmitah*) to evaluate participants' understanding of written texts (Zubaidi 2020). The *insyā'*

component measures participants' ability to synthesize the three preceding language skills in narrative form (Anwar et al. 2023; al-Naqoh 2006). This composition aligns with the objectives of teaching Arabic to non-native speakers, namely, developing the ability to understand, use, and communicate in Arabic accurately through both receptive and productive skills (Ahmad 1979).

Accordingly, this study focused on developing an online Test of Arabic as a Foreign Language (TOAFL) platform based on a Learning Management System (LMS) and constructing an item bank at UIN Syekh Wasil Kediri. The institutional context revealed substantial inefficiency because only one laboratory with 20 computers served as many as 1,500 examinees. To address this constraint, the study applied the 4-D Research and Development model, consisting of Define, Design, Develop, and Disseminate, to establish a Test from Anywhere (TfA) system. The LMS-based platform enables participants to complete the examination flexibly from different locations without attending the Language Center in person, thereby supporting a more efficient and measurable assessment process (Watson, n.d.).

Online testing offers a reliable and efficient response to institutional limitations in facilities and infrastructure, particularly in the context of Industry 4.0, which has intensified the use of internet-based activities (N. Hula et al. 2021). The expanding availability of internet networks across most regions of Indonesia further supports this approach. One survey reported that Indonesians spend an average of 8.5 hours per day accessing the internet (Albantani and Rahmadi 2020).

A Learning Management System (LMS) offers a platform that institutions can optimize for TOAFL administration. An LMS commonly supports online learning activities, ranging from content management to learning evaluation, and it can also integrate formal testing (Rosmiati et al. 2021). Using an LMS as a testing medium can improve assessment efficiency because participants can access the system from different locations whenever an internet connection is available (Yusuf Qathith 2009). An LMS can also accommodate many TOAFL examinees without requiring their physical presence in a classroom. Therefore, this study addressed two main questions: how can an LMS-based online TOAFL platform be developed, and how can the development process be optimized through an expanded item bank and enriched test content to improve TOAFL services? The study consequently proposes a flexible system that allows participants to take the test from any location without attending the UIN Syekh Wasil Kediri Language Center in person.

METHOD

This study employed a qualitative approach within a Research and Development (R&D) framework and followed the 4-D model developed by Thiagarajan et al. The model comprises four stages: Define, Design, Develop, and Disseminate (Thiagarajan et al. 1974). The researchers collected data through interviews, observation, and documentation. They conducted semi-structured interviews with 20 students and 5 lecturers using an interview guide to analyze the need for the TOAFL platform. The informants represented three stakeholder groups: specialists in test media development, Arabic language lecturers, and students as prospective platform users. The researchers used an observation sheet to examine the instructional process and the characteristics of Arabic-language literature relevant to the study. They also collected supporting documents, including curricula, market needs, teaching materials, transcripts, field notes, and other academic records. The researchers analyzed the qualitative data through Miles and

Huberman's interactive model, which involves data reduction, data display, and conclusion drawing and verification. This process provided an empirical foundation for developing test materials that addressed the identified platform requirements (Koentjaraningrat 1994; Moleong 2001; Margono 1997; Black, n.d.; Syalabi 1997).

RESULTS AND DISCUSSION

Needs Analysis and Initial Conditions

Before the development of the online platform, TOAFL administration at UIN Syekh Wasil Kediri involved technical, administrative, and pedagogical problems that directly reduced the effectiveness of Arabic proficiency testing services. Field observations and in-depth interviews with Language Center administrators indicated that the conventional Computer-Based Test (CBT) model relied entirely on the university's computer laboratory and therefore faced substantial limitations. Although CBT generally offers advantages over paper-based testing, particularly automated scoring and faster data processing, its dependence on fixed institutional infrastructure constrained its implementation (Moh. Ismail; Aufa Alfian Musthafa 2018).

The computer laboratory at UIN Syekh Wasil Kediri could accommodate only a limited number of examinees in each session. The Language Center consequently had to organize several testing sessions to serve all participants during a single testing period. Repeated scheduling increased the administrators' technical workload and extended the overall duration of TOAFL administration. This finding supports Ockey's argument that computer-based assessments that depend on institutional hardware remain vulnerable to technical and capacity constraints that can reduce testing efficiency (Ockey 2009).

In addition to facility limitations, the campus-centered CBT model restricted participant flexibility, particularly for students completing off-campus academic programs such as community service or teaching practice. During the interviews, students frequently reported scheduling difficulties, especially when they had to travel considerable distances solely to take TOAFL. Brown emphasizes that sound language assessment should ensure accessibility and participant convenience without compromising measurement validity and reliability. Institutions with large student populations therefore require a more flexible testing model (Brown 2006a).

Language Center administrators also encountered administrative and technical demands, including routine device maintenance, network inspection, and the management of potential equipment failures during testing. Ockey identify this technical complexity as a major weakness of conventional CBT systems that depend on physical infrastructure. These demands increased the burden on human resources and slowed the overall administration process (Ockey 2009).

The needs analysis also identified the importance of a comprehensive and systematically organized item bank. As Alderson explains, an item bank represents a fundamental element of language testing because it supports content validity and maintains a balanced distribution of item difficulty across test administrations. In

the TOAFL context, periodic item renewal and variation in difficulty levels remain essential for maintaining assessment quality (John Charles Alderson 2010a).

Interviews with 20 students and 5 lecturers showed a shared preference for a Learning Management System (LMS) as the most appropriate solution. Participants regarded the LMS as stable, affordable, accessible, and familiar within the academic environment of UIN Syekh Wasil Kediri. Kasabova argues that an LMS provides considerable flexibility for learning and assessment and supports multiple media formats, including audio for the *istimā* section (Kasabova et al. 2023). Sarwadi further explains that an LMS can automate scoring, facilitate item-bank integration, and support efficient administration of language proficiency examinations (Sarwadi et al. 2025). Moreover, LMS security functions, including item randomization, time limits, and participant activity monitoring, can strengthen academic integrity. This finding corresponds with Hamp-Lyons's view that a reliable language assessment must ensure fairness and objectivity while minimizing opportunities to manipulate test results (Hamp-Lyons 2017).

Collectively, these conditions demonstrate an urgent need to innovate TOAFL administration through a more flexible, efficient, and adaptive system. The needs analysis provides a strong empirical and theoretical basis for developing an LMS-based TOAFL platform that can improve the quality, scalability, and accessibility of Arabic language proficiency testing services at UIN Syekh Wasil Kediri.

Development of the TOAFL Item Bank and Test Materials

Development of the Test Blueprint and Specification Table

The development of the TOAFL item bank began with the preparation of a test blueprint and specification table as the principal foundation for test quality and validity. This stage played a strategic role because the blueprint mapped the relationship among the measurement objectives, the competencies under assessment, and the forms and characteristics of the items presented to participants. Without a clear and systematic blueprint, an item bank may lack consistency and fail to represent the intended language proficiency construct.

The researchers developed the TOAFL blueprint with reference to the standards used in Islamic higher education institutions, particularly standards that function as benchmarks for students' Arabic proficiency. They considered the proportion of items in each test component, the variety of texts and audio recordings, item difficulty, and the competencies under measurement. They distributed the items to ensure that istimā', qirā'ah, qawā'id, and *mufradāt* each contributed proportionately to a comprehensive representation of participants' Arabic proficiency (Sa'idah et al. 2022).

Figure 1. TOAFL Blueprint Document

BLUEPRINT DAN KISI-KISI TES TOAFL
(Test of Arabic as a Foreign Language)

Dokumen blueprint TOAFL ini disusun sebagai acuan resmi dalam pengembangan bank soal dan pelaksanaan Tes TOAFL berbasis Learning Management System (LMS) di lingkungan Perguruan Tinggi Keagamaan Islam. Blueprint ini berfungsi untuk menjaga keselarasan antara tujuan tes, kompetensi yang diukur, serta karakteristik dan proporsi butir soal.

A. Tujuan Tes TOAFL

Tes TOAFL bertujuan untuk mengukur kemampuan bahasa Arab mahasiswa sebagai bahasa asing secara komprehensif, yang meliputi kemampuan menyimak (*istimâ'*), membaca (*qirâ'ah*), penguasaan kaidah bahasa (*qawâ'id*), serta kemampuan menganalisis kosakata (*mufradât*). Hasil tes digunakan sebagai dasar evaluasi kemampuan bahasa Arab akademik mahasiswa.

B. Komponen Tes, Jumlah Soal, dan Bobot Penilaian

No.	Komponen Tes	Kompetensi yang Diukur	Jumlah Soal	Bobot Nilai
1	*Istimâ'*	Pemahaman bahasa Arab lisan (makna umum, detail, inferensi)	30	25%
2	*Qirâ'ah*	Pemahaman teks tertulis (ide pokok, detail, inferensi)	50	25%
3	*Qawâ'id*	Penguasaan struktur dan kaidah bahasa Arab	15	25%
4	*Mufradât*	Kemampuan menyusun *mufradât* menjadi sebuah kalimat efektif	5	25%

Berdasarkan komposisi tersebut, jumlah keseluruhan butir soal TOAFL adalah sebanyak 100 soal dengan pembobotan nilai yang seimbang antar komponen. Pembagian bobot yang sama dimaksudkan untuk memberikan penilaian yang proporsional terhadap setiap aspek kemampuan bahasa Arab yang diukur.

C. Distribusi Tingkat Kesukaran Soal

Distribusi tingkat kesukaran soal disusun secara proporsional agar tes mampu membedakan kemampuan peserta secara objektif.

Kategori Tingkat Kesukaran	Persentase	Keterangan
Mudah	30%	Mengukur penguasaan dasar
Sedang	50%	Mengukur kemampuan inti
Sulit	20%	Mengukur kemampuan lanjut

D. Bentuk dan Karakteristik Soal

Bentuk soal TOAFL adalah pilihan ganda untuk komponen *istimâ'*, *qirâ'ah*, *qawâ'id*, dan *mufradât*. Seluruh soal disajikan secara daring melalui LMS dengan fitur randomisasi soal, pembatasan waktu, dan satu kali kesempatan pengerjaan.

E. Standar Teknis Pelaksanaan Tes

Tes TOAFL dilaksanakan secara daring melalui LMS resmi institusi. Sistem secara otomatis mencatat aktivitas peserta melalui log LMS, melakukan penilaian, serta menghasilkan rekap nilai secara otomatis.

The blueprint also addressed the diversity of texts and recordings. For the *istimā'* component, the researchers designed varied audio materials that included short dialogues, monologues, and simple discourse delivered at an appropriate speed and intonation for a proficiency test. For the *qirā'ah* component, they used narrative, descriptive, and informative texts. This variation reduced item monotony and provided a broader representation of participants' ability to comprehend different types of Arabic discourse. Alderson emphasizes that varied linguistic input constitutes an important requirement for a proficiency test because it enables the instrument to represent actual language ability more accurately (John Charles Alderson 2010a).

The blueprint further organized items into progressive difficulty levels, ranging from easy to moderate and difficult. This distribution enabled the test to differentiate participants' proficiency more accurately. Brown argues that an effective proficiency test must measure a broad ability spectrum rather than assess only basic competence. The researchers therefore distributed item difficulty proportionately to produce a more objective representation of participant proficiency (Brown 2006a).

The blueprint also specified the competencies measured in each component. The *istimā'* component assessed comprehension of general meaning, detailed information, and simple inferences from audio recordings. The *qirā'ah* component assessed comprehension of main ideas, explicit and implicit information, and vocabulary meaning in context. The *qawā'id* and *mufradāt* components focused on language structures, grammatical patterns, and functional vocabulary commonly used in academic contexts. This alignment between competencies and item formats reflects the principle of content validity, namely the correspondence between test content and the construct under measurement (John Charles Alderson 2010a).

After completing the blueprint, the researchers developed a detailed specification table. The table translated the blueprint into technical indicators, item formats, the number of items assigned to each indicator, and scoring criteria. A detailed specification table helped the item writers maintain consistency across items and facilitated subsequent evaluation and revision. Alderson identifies clear test specifications as an important instrument for maintaining consistency, particularly when institutions reuse an item bank across multiple testing periods (John Charles Alderson 2010a).

Figure 2. TOAFL Test Specification and Item Instrument

Instrumen Soal TOAFL

A. Pedoman Soal TOAFL per Level Kemampuan

Komponen	Indikator	Level	Jumlah Soal	Bentuk Soal
Istimā'	Memahami dialog sederhana	Mubtadi'	10	PG
Istimā'	Memahami ide pokok	Mutawassith	10	PG
Istimā'	Menarik kesimpulan	Lanjutan	10	PG
Qirā'ah	Makna kosakata	Mubtadi'	15	PG
Qirā'ah	Ide pokok teks	Mutawassith	20	PG
Qirā'ah	Analisis isi teks	Lanjutan	15	PG
Qawā'id	Jenis kata	Mubtadi'	5	PG
Qawā'id	Struktur kalimat	Mutawassith	5	PG
Qawā'id	Ketepatan tarkib	Lanjutan	5	PG
Mufradāt	Ekspresi tertulis terkontrol	Semua level	5	PG

B. Contoh Soal TOAFL dan Opsi Jawaban

[الاستماع]

1. ما موضوع الحوار؟
a. الدراسة
b. السفر
c. العمل
d. الأسرة
الإجابة الصحيحة : a

2. من المتحدث في الحوار؟
a. طالب
b. معلم
c. موظف
d. طبيب
الإجابة الصحيحة : b

3. ما هدف الحوار؟
a. طلب معلومات
b. تقديم شكوى
c. إعطاء تعليمات
d. الدعوة
الإجابة الصحيحة : a

[القراءة]

1. ما الفكرة الرئيسية للنص؟
a. أهمية العلم
b. وقت الفراغ
c. الأسرة
d. الصحة
الإجابة الصحيحة: a

2. ما معنى كلمة(يجتهد)؟
a. يعمل بجد
b. ينام
c. يلعب
d. يسافر
الإجابة الصحيحة : a

3. ما الاستنتاج الصحيح من النص؟
a. العلم أساس النجاح
b. المال أهم
c. الصحة ثانوية
d. العمل غير مهم
الإجابة الصحيحة: a

Through the systematic preparation of the blueprint and specification table, the TOAFL item bank acquired a strong conceptual and technical foundation. This stage ensured that all developed items met academic standards and aligned to measure Arabic proficiency as a foreign language.

Item Development

Following the preparation of the blueprint and specification table, the researchers systematically wrote and developed the TOAFL items. They followed the established blueprint closely so that each item represented the Arabic competence under measurement. At this stage, they developed an item bank for four principal TOAFL components, namely *istimā', qirā'ah, qawā'id,* and *insyā'*, while considering content validity, variation in difficulty, and clarity of item construction (Maul and Rowin 2026).

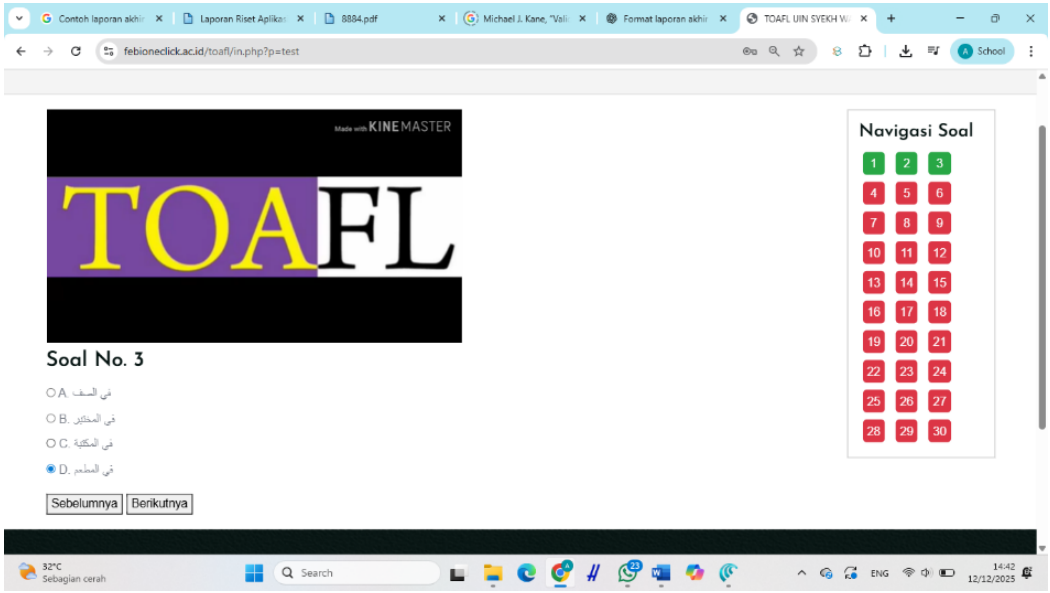
1. Development of Istimā' Items

The development of *istimā'* items focused on participants' ability to comprehend spoken Arabic as a foreign language. The researchers presented the items through audio materials that included conversations, monologues, and short dialogues. They considered pronunciation clarity, natural

intonation, and a speaking rate appropriate to participant proficiency. These varied recordings enabled the test to measure listening comprehension across different formal and semi-formal communicative contexts.

The researchers organized the audio recordings from lower to higher difficulty levels to differentiate participants' listening proficiency accurately. Buck argues that a sound listening test should present authentic and varied audio input to support valid measurement of spoken-language comprehension. The *istimā'* items therefore assessed comprehension of general meaning, specific information, and simple inferences based on the material participants heard (Buck 2001).

Figure 3. Sample *Istimā'* Audio and Items on the LMS

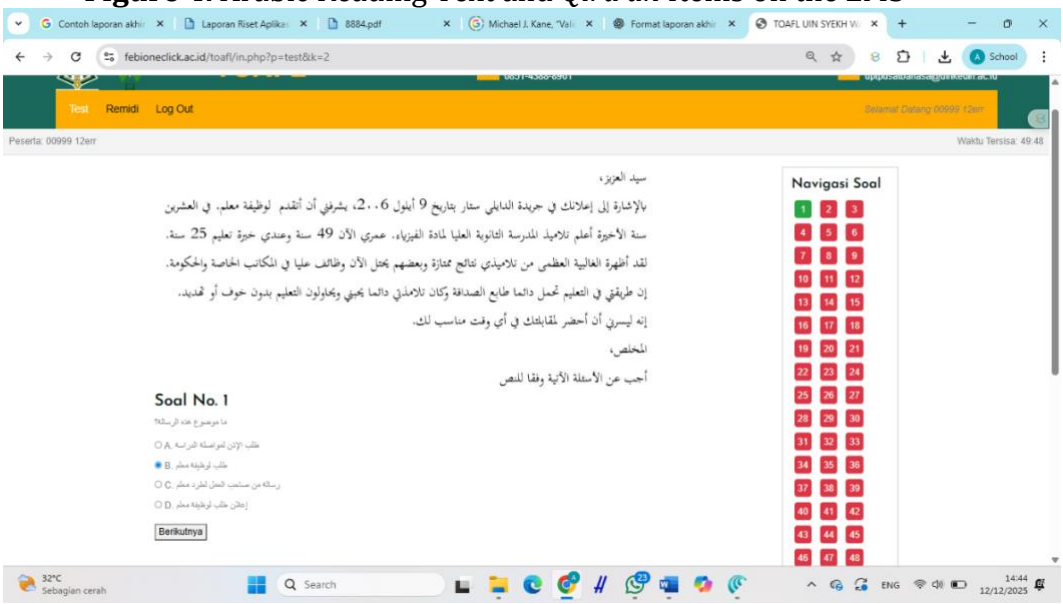


2. Development of *Qirā'ah* Items

The development of *qirā'ah* items focused on measuring participants' comprehension of written Arabic texts. The researchers used narrative, descriptive, and argumentative discourse of varying lengths to provide a comprehensive representation of reading ability in academic and non-academic contexts.

Each reading text included questions that assessed meaning comprehension, identification of the main idea, understanding of explicit and implicit information, and the ability to draw simple inferences. Alderson stresses that effective reading tests should assess more than word recognition because they must also measure the ability to understand meaning and relationships among ideas. Accordingly, the *qirā'ah* items assessed both literal comprehension and participants' analytical engagement with Arabic texts (J. Charles Alderson 2000).

Figure 4. Arabic Reading Text and *Qirā'ah* Items on the LMS

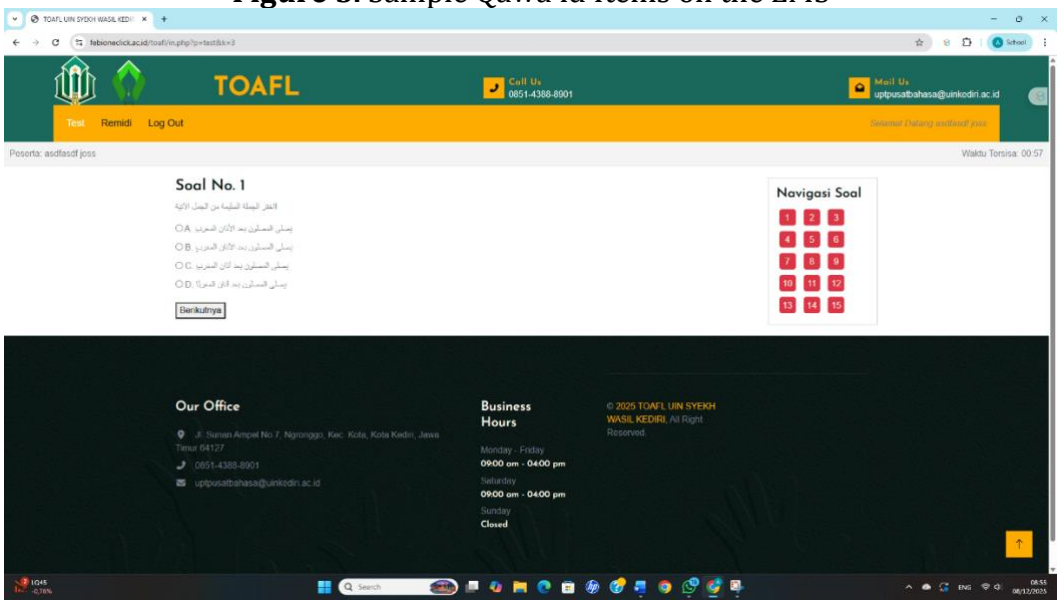


3. Development of *Qawā'id* Items

The development of qawā'id items focused on measuring participants' mastery of Arabic language structures. The items covered sentence-structure analysis, understanding of *i'rāb*, grammatical meaning, and identification of common errors made by learners of Arabic as a foreign language. The researchers varied the item formats by asking participants to complete sentences, select appropriate structures, and identify grammatical errors.

The development process recognized grammar as an important component of language competence while requiring its functional and contextual assessment. Brown argues that grammar assessment should remain connected to language use so that the resulting measurement becomes more meaningful. The TOAFL item bank therefore assessed *qawā'id* through contextualized items rather than isolated rule memorization (Brown 2006b).

Figure 5. Sample Qawā'id Items on the LMS



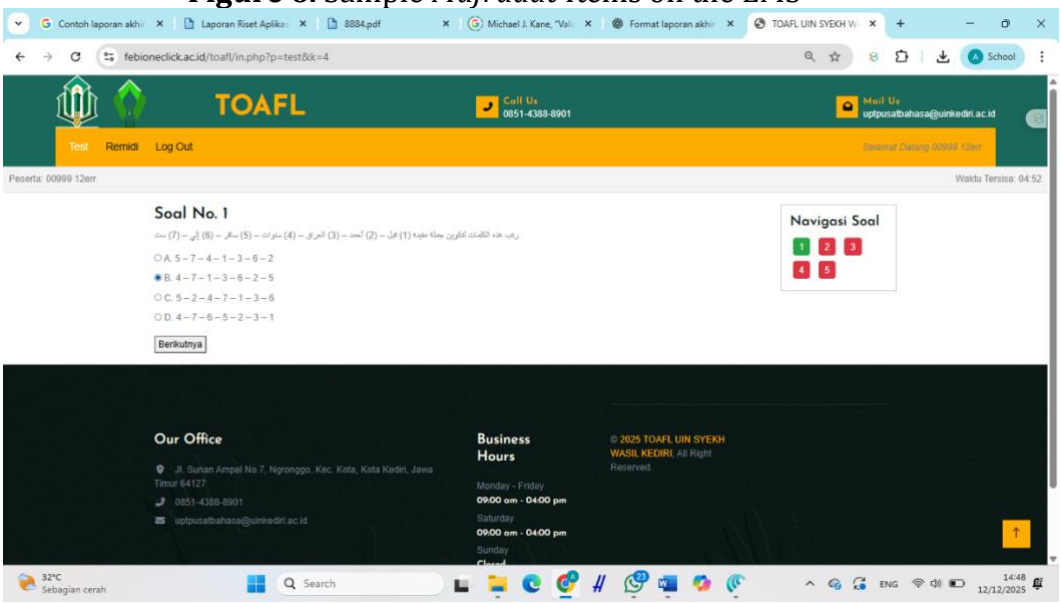
4. Development of *Mufradāt* Items

The development of *mufradāt* items focused on measuring participants' contextual mastery of Arabic vocabulary. The researchers designed the items to assess participants' understanding of vocabulary meaning in sentences or discourse and their recognition of relevant synonyms and antonyms. They selected this contextual approach because vocabulary mastery cannot be separated from its communicative use.

Nation explains that vocabulary mastery involves understanding a word's meaning, form, and appropriate contextual use. The *mufradāt* items in the TOAFL item bank therefore assessed more than isolated lexical meaning because they presented vocabulary in contexts that approximated authentic Arabic use (Nation 2006).

After developing all items for the four TOAFL components, the researchers conducted internal editing and format verification. They reviewed linguistic appropriateness, instructional clarity, format consistency, and technical compatibility with the LMS platform. This stage minimized non-substantive errors that could affect test quality and participant experience. Through systematic development and editing, the resulting TOAFL item bank achieved an adequate academic standard and became ready for product trials.

Figure 6. Sample *Mufradāt* Items on the LMS



Item Quality Analysis

Alderson explains that the quality of language assessment depends substantially on the quality of the item bank and the consistency of difficulty across items. The researchers therefore developed the item bank with reference to the TOAFL blueprint and a planned distribution of difficulty levels (John Charles Alderson 2010a).

Expert validation constituted an important stage of this development study because it determined whether the product met academic, pedagogical, and technical standards before broader implementation. Two groups of validators evaluated the LMS-based online TOAFL platform and its integrated item bank: subject-matter experts and media experts. Their involvement ensured that the

product demonstrated appropriate test content and substance while also meeting the technical and interface requirements needed for accurate and convenient test administration.

After developing and internally editing the items for the four TOAFL components, namely *istimā', qirā'ah, qawā'id,* and *insyā'*, the researchers analyzed item quality. This analysis examined whether each item functioned effectively as a measure of Arabic proficiency and met the standards of a sound proficiency test. The initial analysis focused on two principal indicators: item difficulty and item discrimination.

The item difficulty analysis determined the proportion of examinees who could answer each item correctly. Proficiency tests generally require a substantial proportion of moderately difficult items because such items can differentiate participants across different levels of language ability. The results indicated that most items fell within moderate and acceptable difficulty categories. The developed item bank therefore generally reflected the characteristics of an Arabic proficiency test. This result supports Alderson's argument that a balanced difficulty distribution constitutes an important prerequisite for the validity and reliability of proficiency testing (John Charles Alderson 2010b).

The researchers also analyzed item discrimination, namely an item's capacity to distinguish participants with higher proficiency from those with lower proficiency. A strong discrimination index indicates that an item measures ability accurately rather than producing random results. The analysis showed that most items had adequate discrimination and could function effectively as measures of Arabic proficiency. Brown emphasizes that items with low discrimination provide limited information and require revision or removal to prevent them from weakening overall test quality (Brown 2006b).

Nevertheless, the analysis identified a small number of items with excessively low or high difficulty. Items that were too easy provided limited differentiation among participants, whereas items that were too difficult risked measuring factors unrelated to language ability, including guessing and fatigue. The researchers did not immediately remove these items. Instead, they revised the wording, adjusted the context, or improved the response options to achieve more proportionate difficulty. This revision process follows Alderson's principle that item analysis should lead to systematic improvement of an item bank (John Charles Alderson 2010b).

Item Difficulty Analysis

The item difficulty analysis determined the extent to which participants could answer each test item correctly. Moderately difficult items generally provide the greatest value in proficiency testing because they can distinguish participants across different levels of language ability. Based on expert review of all TOAFL items, the distribution of difficulty across the test components appeared relatively proportionate.

In the *istimā'* component, 30% of the 30 items were classified as easy, 50% as moderate, and 20% as difficult. The researchers calculated item difficulty using Alderson's classical test theory framework with the following formula: $\mathrm{FV} = \frac{B}{N}$. In this formula, FV represents item difficulty, B represents the number of participants who answered the item correctly, and N represents the total number of test participants. This distribution indicates that the *istimā'* items measured listening proficiency progressively, from basic comprehension to the understanding of more

complex information. A similar pattern appeared in the *qirā'ah* component, in which most of the 50 items fell within the moderate category and the easy and difficult items remained balanced. The reading items therefore accommodated variation in participants' ability to comprehend Arabic texts.

The *qawā'id* component contained fewer items, but its difficulty distribution still demonstrated a relatively balanced pattern. The items represented both basic and advanced competence in Arabic structures and grammatical rules. The *insyā'* component consisted of five limited-response items, most of which fell within the moderate category. This level was considered appropriate because the test should measure basic writing as a productive skill without imposing excessive difficulty that could restrict participants' expression of their language ability.

Overall, the item difficulty analysis indicated that most TOAFL items fell within the moderate category, which corresponds to the characteristics of a proficiency test. The researchers recommended revising several items that appeared excessively easy or difficult to maintain a proportionate distribution and protect the quality of Arabic proficiency measurement.

Item Discrimination Analysis

In addition to difficulty, the item quality analysis examined item discrimination. Item discrimination indicates how effectively an item distinguishes participants with higher ability from those with lower ability. Items with strong discrimination provide more accurate information about participant proficiency.

Following H. Douglas Brown, item discrimination (ID) refers to an item's ability to distinguish higher-performing test takers from lower-performing test takers. A sound item should be answered correctly more frequently by the higher-performing group than by the lower-performing group using the formula $ID = \frac{U-L}{\frac{1}{2}(U+L)}$ or, equivalently, $= \frac{U-L}{n}$. In these formulas, ID denotes item discrimination, U denotes the number of correct responses in the upper group, L denotes the number of correct responses in the lower group, and n denotes the number of participants in either group because the two groups have equal size (Brown 2006b). The item discrimination summary showed that most TOAFL items fell within the good to very good categories. Of the 100 items analyzed, the majority demonstrated adequate discrimination and warranted retention. These items effectively measured participants' Arabic proficiency and corresponded to the objectives of proficiency testing.

A small number of items, however, demonstrated fair or poor discrimination. These items generally contained unclear wording, contexts that were either overly simple or excessively complex, or response options that did not function effectively as distractors. Rather than removing the items immediately, the researchers recommended revising their wording, adjusting their contexts, or strengthening their distractors. This approach maintained the breadth of the item bank while allowing the revised items to be reused after improvement.

Conclusion of the Item Quality Analysis

The analysis of difficulty and discrimination, supported by expert review, indicated that the developed TOAFL items demonstrated good quality and were suitable for use as an Arabic language proficiency instrument. The proportionate

distribution of item difficulty and the predominance of items with good discrimination showed that the TOAFL item bank met fundamental principles of language test development.

The analysis also provided an empirical basis for continuous improvement of the item bank, particularly through revision of items with extreme difficulty or low discrimination. Item quality analysis therefore functioned not only as an evaluative procedure but also as a foundation for the sustainable development of LMS-based TOAFL administration in Islamic higher education institutions (Al-Fraihat et al. 2020).

Figure 7. Item Difficulty Distribution and Item Discrimination Summary

ANALISIS KUALITAS SOAL TOAFL

Dokumen ini menyajikan hasil analisis kualitas soal TOAFL berdasarkan telaah ahli (expert judgment) terhadap dokumen soal TOAFL 2025.

A. Tim Ahli Analisis Soal

1. H. Saifullah, M.Ag. – Kepala UPT Bahasa UIN Syekh Wasil Kediri
2. Eko Budi Hartanto, M.Pd.I. – Dosen TOAFL di Prodi Bahasa Arab UIN Syekh Wasil Kediri

B. Tabel Analisis Tingkat Kesukaran Soal

Komponen	Jumlah Soal	Mudah	Sedang	Sulit
Istima'	30	9 (30%)	15 (50%)	6 (20%)
Qira'ah	50	15 (30%)	25 (50%)	10 (20%)
Qawaid	15	5 (33%)	7 (47%)	3 (20%)
Insya'	5	1 (20%)	3 (60%)	1 (20%)

C. Rekap Daya Pembeda Butir Soal

Kategori Daya Pembeda	Rentang Nilai	Jumlah Soal	Keterangan
Sangat Baik	> 0,40	42	Dipertahankan
Baik	0,30 – 0,39	38	Dipertahankan
Cukup	0,20 – 0,29	15	Direvisi
Kurang	< 0,20	5	Direvisi/Diganti

D. Kesimpulan Analisis

Secara umum, kualitas soal TOAFL menunjukkan tingkat kesukaran yang proporsional dan daya pembeda yang baik. Mayoritas butir soal layak digunakan, sementara sebagian kecil direkomendasikan untuk revisi.

Availability of the Final Item Bank

The final stage of item-bank development established a collection of items ready for operational test administration. After completing expert validation, product trials, quality analysis, and empirical revision, the researchers compiled a final item bank containing enough items to generate several TOAFL test forms. Multiple forms maintain item variation and reduce excessive item repetition across successive testing periods.

The final item bank maintained a balance among test components, represented different difficulty levels, and covered the Arabic competencies under measurement. This structure gives the Language Center flexibility to construct TOAFL forms for different testing periods without writing an entirely new set of items each time. Brown identifies a rich and systematically organized item bank as an indicator of a mature language assessment system in higher education (Brown 2006b).

In addition to supporting regular TOAFL administration, the final item bank creates opportunities for future adaptive testing. Because the LMS contains items whose quality has been analyzed, future developers can design a system that progressively adjusts item difficulty to participant ability. This possibility corresponds to contemporary developments in language assessment, which increasingly use technology to improve the accuracy and efficiency of proficiency measurement (Kane 2013). The final item bank therefore represents both an important research output and a strategic foundation for developing more innovative, sustainable, and responsive TOAFL services for institutions and examinees.

Development Process of the LMS-Based Online TOAFL Platform

The development of the LMS-based online TOAFL platform followed the 4-D Research and Development model, which comprises Define, Design, Develop, and Disseminate. This section describes three principal stages that the researchers implemented systematically to produce an online testing prototype suitable for assessing Arabic language proficiency at UIN Kediri. The discussion presents the logical development process from needs analysis to product trials.

After identifying the requirements, the researchers designed the LMS-based online TOAFL platform by considering user flow, interface design, test security, and item-bank integration. They structured the participant and administrator workflows to promote clarity and efficiency. They also prioritized readability, clear navigation, and visual simplicity in the interface.

1. Define Stage: Needs Analysis and Initial Product Specifications

The Define stage established the initial product specifications and therefore formed the foundation of the development process. The researchers conducted a needs analysis through interviews, field observations, and a review of documents related to TOAFL administration standards. They interviewed Language Center administrators, Arabic language lecturers, and students as prospective platform users. They also directly observed the existing CBT-based TOAFL process to identify its technical and administrative constraints.

The needs analysis showed that the previous TOAFL system faced substantial limitations in administrative flexibility, technical efficiency, and score processing. Language assessment theory similarly requires modern assessment systems to ensure accessibility, reliability, and technical feasibility. The LMS-based TOAFL platform therefore needed to accommodate different test components, including istimā', qirā'ah, qawā'id, and mufradāt, in accordance with the standards used in Islamic higher education institutions (Brown 2006b).

The literature review indicated that an effective LMS-based online testing system should present audio items reliably, provide clear navigation,

and support automated scoring. Sarwadi states that LMS-based language assessment can integrate multimedia functions and automated score processing, thereby increasing administrative efficiency. The system also required security features that could reduce opportunities for misconduct, including item randomization and time limits. Brown similarly argues that valid and reliable assessment requires a secure testing environment that minimizes academic dishonesty (Sarwadi et al. 2025).

The needs analysis further showed that Language Center administrators should be able to manage the platform without advanced technical expertise. This requirement corresponds with Abdulquadir's view that an effective LMS-based evaluation system should be easy for instructors and administrators to operate and should not require complex technical support (Abdulquadir et al. 2023).

The Define stage produced the following basic product specifications:

a. the platform must accommodate four principal TOAFL components.
b. the platform must support reliable audio playback.
c. the platform must provide item and response-option randomization.
d. the platform must generate automated reports; and
e. administrators must be able to manage the platform without specialized technical competence.

These specifications provided the foundation for designing the online TOAFL platform in the subsequent stage.

2. Design Stage: Conceptual and Technical Design

The Design stage focused on the platform's user flow, navigation structure, user interface, and technical LMS configuration. The researchers developed a functional blueprint, simulated the proposed workflow, and created interface mockups for internal testing.

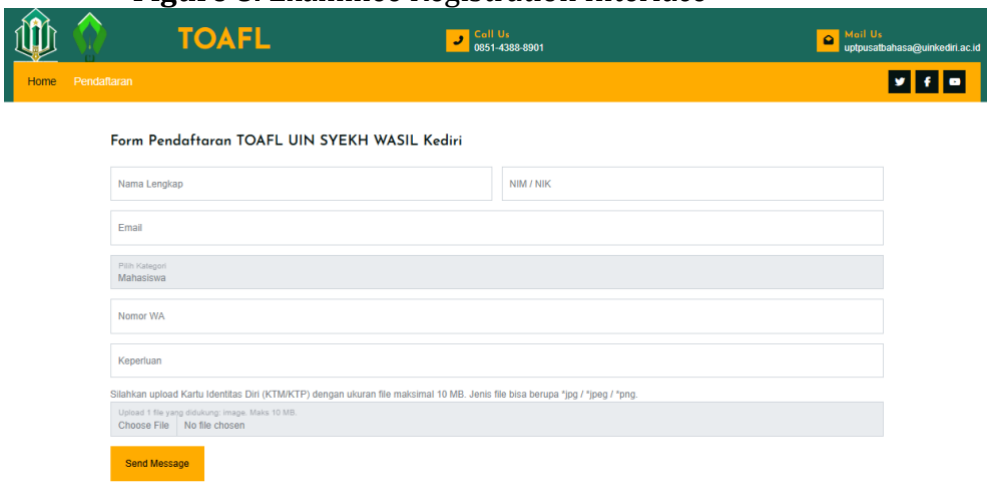
a. Examinee Workflow Design

The researchers designed the examinee workflow to support a clear, systematic, and accessible testing process. The workflow included:

- account registration and verification.
- access to the test instruction page.
- navigation to each test component.
- completion of the test items; and
- automatic score generation.

The workflow design followed Alderson's principle that test takers require clear and accessible instructions to reduce construct-irrelevant variance. The researchers therefore simplified navigation at every stage and provided instructions that participants could understand easily (John Charles Alderson 2010b).

Figure 8. Examinee Registration Interface



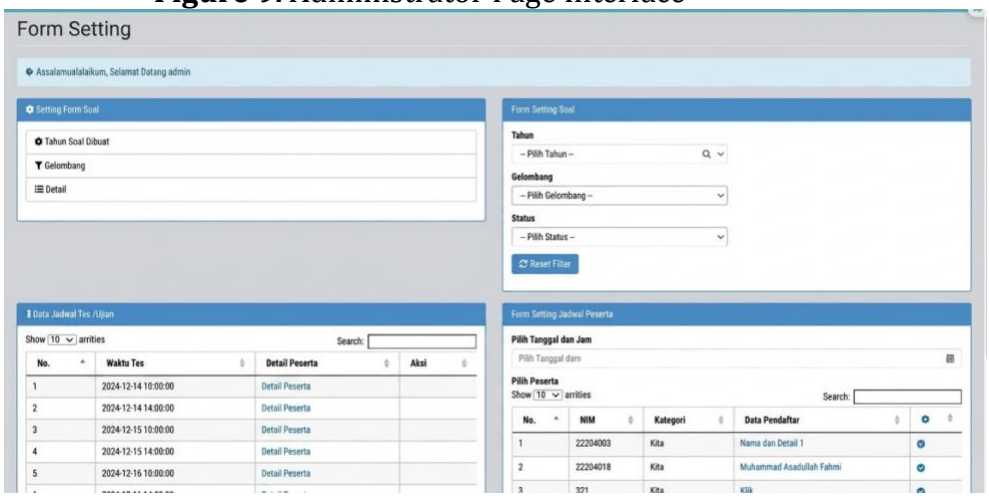
b. Administrator Workflow Design for the Language Center

The researchers designed the administrator workflow to support effective and efficient test management. The available administrative functions included:

- test schedule configuration.
- uploading test items and audio files.
- configuring test duration and regulations.
- monitoring participant activity in real time; and
- automatically downloading test scores.

This design reflects Sommerville's argument that an LMS-based assessment system should enable administrators to organize content and monitor test administration easily (Sommerville 2016).

Figure 9. Administrator Page Interface

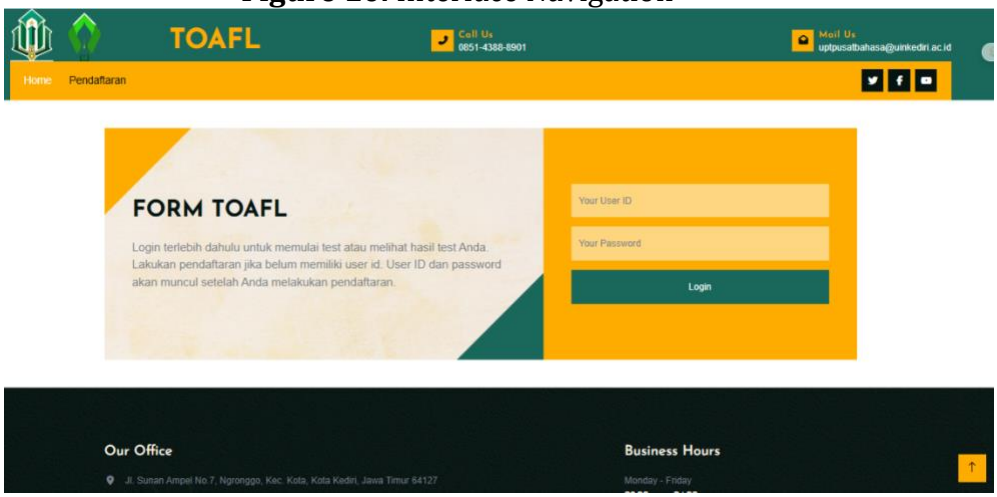


c. Interface and Navigation Design

The interface design emphasized readability, visual simplicity, and layout consistency. The researchers placed each TOAFL component within a clear and accessible menu structure. They also arranged fonts, white space, and interface elements to help participants focus on the test items without unnecessary confusion.

Hamp-Lyons argues that interface quality can influence test-taker performance, particularly in computer-based assessments that require sustained interaction with digital media. The researchers therefore conducted internal prototype testing to ensure that the TOAFL interface addressed user requirements (Hamp-Lyons 2017).

Figure 10. Interface Navigation



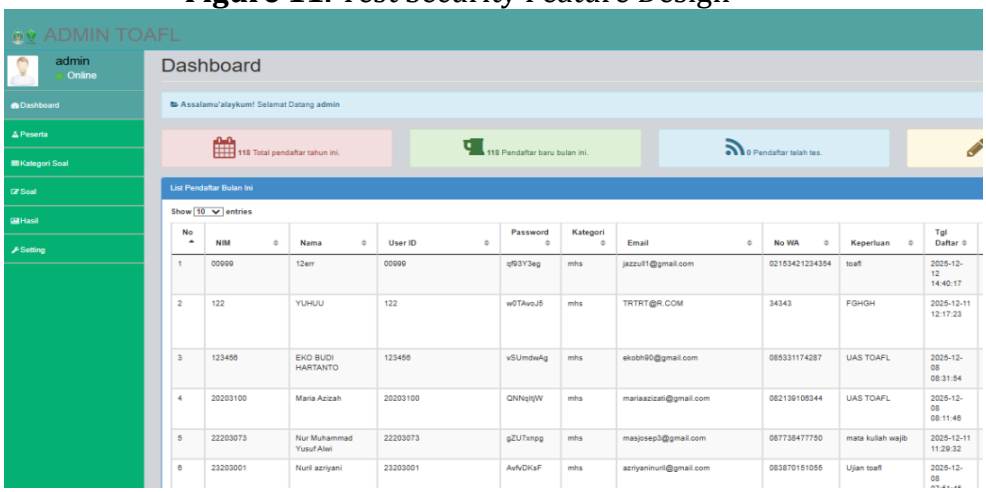
d. Test Security Feature Design

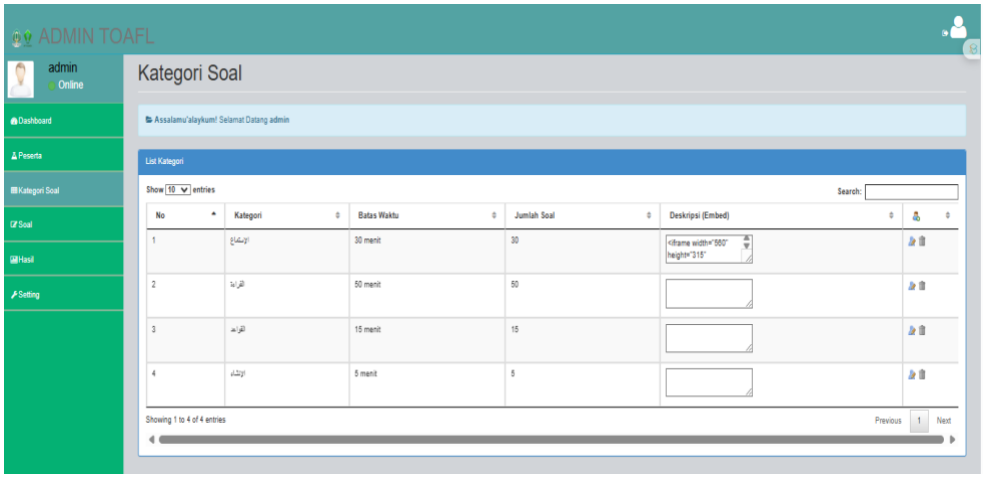
The researchers designed security controls to protect academic integrity. These controls included:

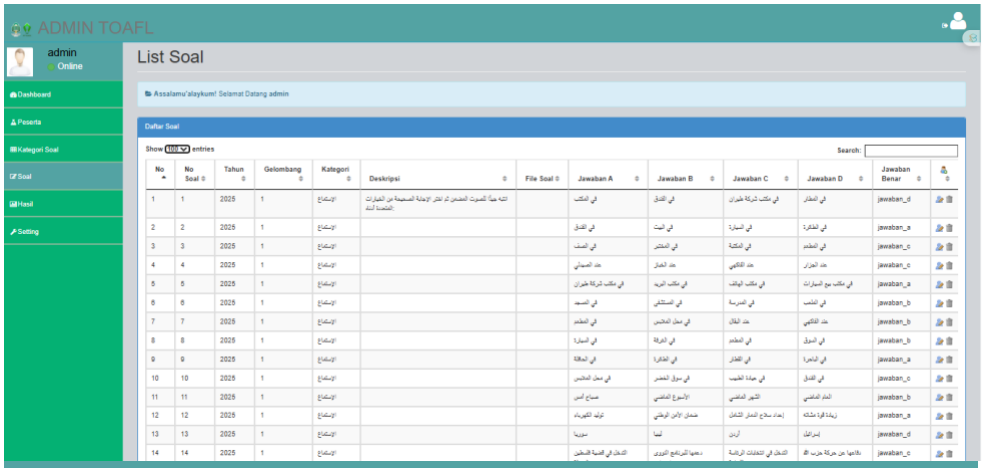
- test time limits.
- restrictions on backward navigation when required.
- randomization of items and response options; and
- a single permitted test attempt.

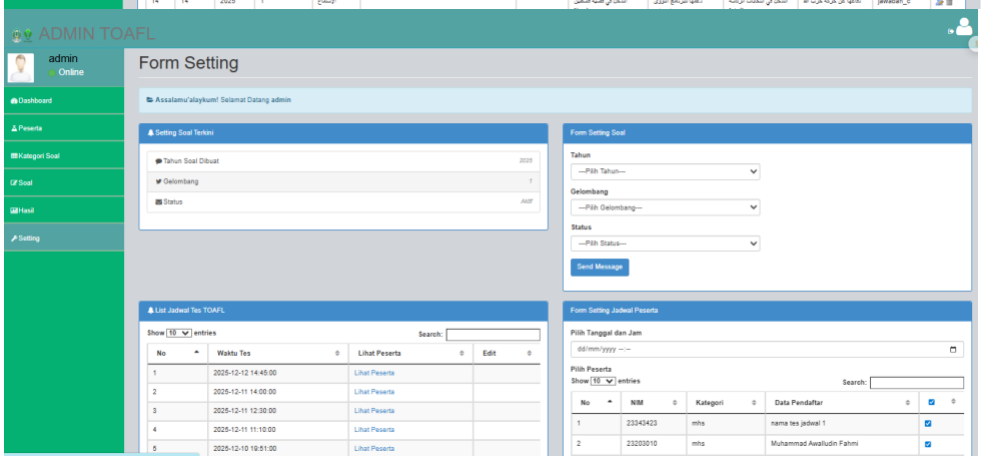
These functions remain important because, as Brown states, secure testing protocols support the reliability of high-stakes language assessment (Brown 2006b). The Design stage concluded with the development of mockups and an internal trial simulation before implementation in the LMS.

Figure 11. Test Security Feature Design









3. Disseminate Stage: Platform User Trials

Limited Trial

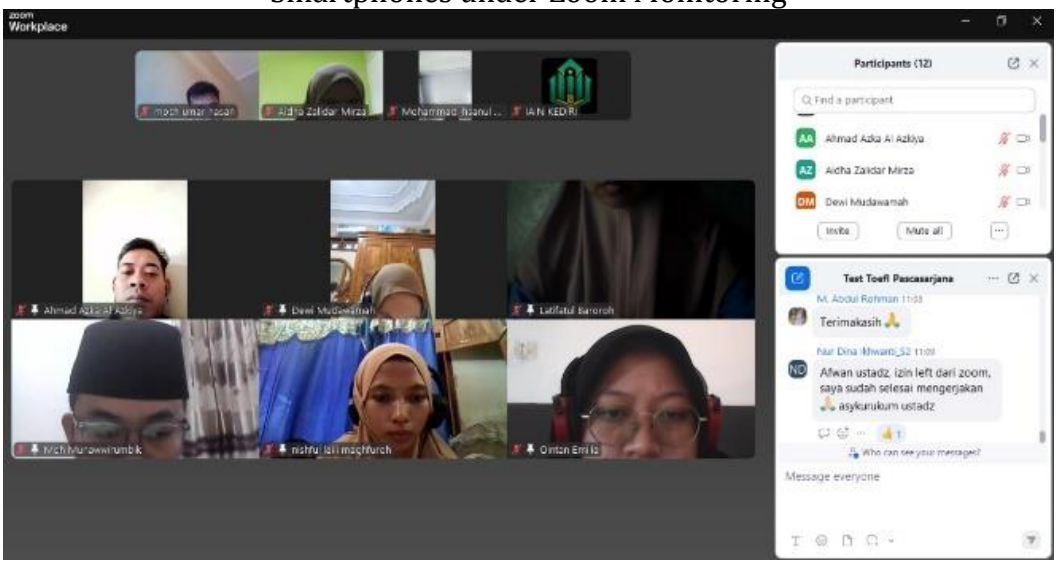
The limited trial constituted the initial product implementation stage and examined the technical stability and basic functionality of the LMS-based online TOAFL platform before broader use. The trial involved 20 students selected to represent the characteristics of TOAFL examinees at UIN Syekh Wasil Kediri. During this stage, the researchers focused on the platform's technical operation, access performance, and participants' initial experience of interacting with the online testing system.

The limited-trial results showed that participants could access the TOAFL platform through different devices, including laptops and smartphones, without substantial technical problems. This cross-device accessibility indicated adequate platform compatibility for participants with varying device conditions. The istimā' audio function also operated reliably without noticeable delay or playback disruption, enabling participants to listen to the audio materials clearly and continuously.

Regarding the interface, most participants reported that the platform provided clear, systematic, and comprehensible navigation. The menu structure and test-page layout enabled participants to move between sections without losing focus on item content. Clear navigation improved participant comfort and reduced technical errors unrelated to the language competence under measurement.

User responses also indicated a more positive perception of testing flexibility. Most participants considered the online TOAFL platform more flexible and practical than conventional CBT, which required physical attendance in a computer laboratory. They could access the test from different locations without strict dependence on institutional space and scheduling, while the platform maintained the quality of test administration.

Figure 12. Participants Accessing the Test through Laptops and Smartphones under Zoom Monitoring

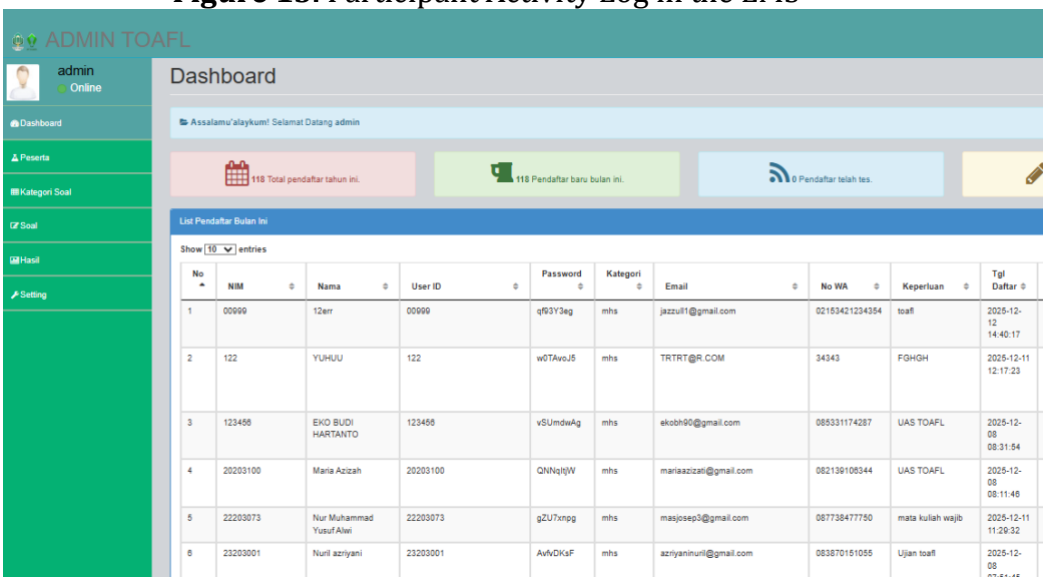


The limited-trial finds that LMS-based online testing offers greater flexibility and convenience than conventional testing, particularly in higher education contexts with large participant populations and high mobility requirements (Tahir et al. 2025; Herniawati et al. 2025; Cao 2023). The results therefore indicated that the LMS-based TOAFL platform met the requirements for technical feasibility and user convenience, justifying its progression to a larger-scale trial in accordance with standard R&D evaluation phases (Branch 2009).

Large-Scale Trial

The large-scale trial represented the next stage of the LMS-based online TOAFL platform development process and examined system readiness under conditions approximating actual test administration. The researchers increased the number of participants to 100 students and simulated the standard TOAFL operating procedures, including test duration, time configuration, and supervision by Language Center administrators. The trial therefore provided a more realistic representation of platform performance during regular use.

Figure 13. Participant Activity Log in the LMS



The large-scale trial showed that the platform remained stable during simultaneous access by many participants. The researchers observed no significant technical disruptions, such as delayed page loading, audio playback failure, or server downtime. This stability indicated that the platform had sufficient technical capacity to support large-scale TOAFL administration in a higher education environment.

The large-scale trial also showed that the LMS activity log recorded the complete testing process accurately. The system automatically documented participant login times, test duration, movement between items, and test completion. These records supported both administrative documentation and the transparency and accountability of assessment. Language Center administrators could also use the log data for monitoring, technical evaluation, and investigation of participant complaints or operational problems.

Automated score processing represented another prominent result of the large-scale trial. After participants completed the test, the system generated and summarized their scores without manual processing. This automation accelerated Language Center services, reduced the risk of human error, and improved administrative efficiency. The system also enabled the

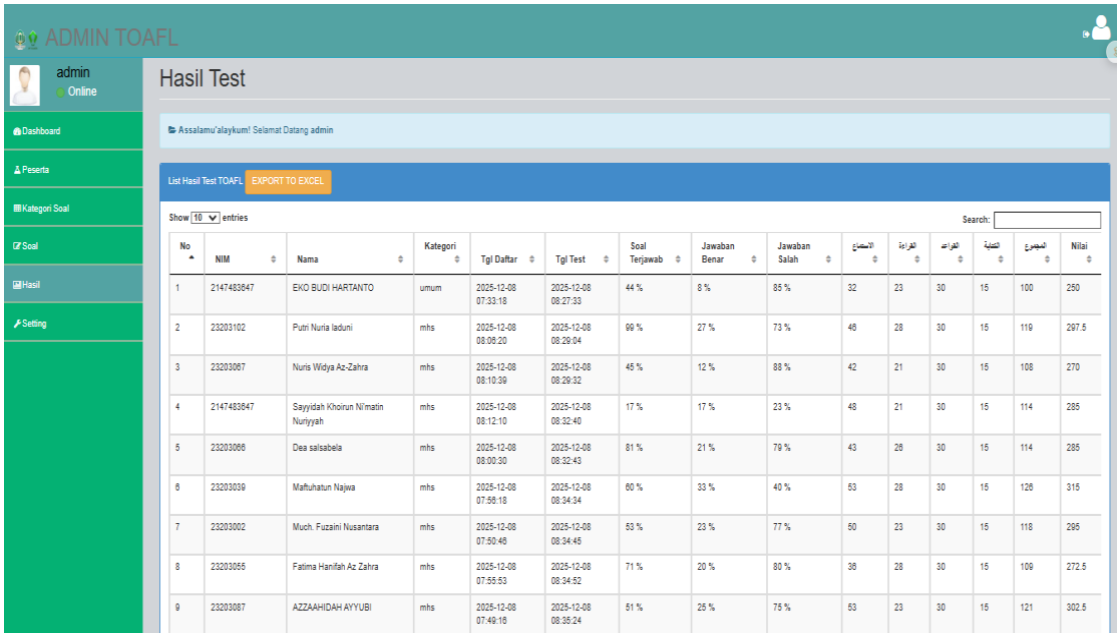
institution to communicate TOAFL results more rapidly than the conventional CBT model, which required additional processing stages.

Overall, the large-scale trial provided empirical evidence that the developed LMS-based online TOAFL platform met the requirements for stability, accuracy, and service efficiency. These findings indicated that the platform was ready for regular TOAFL administration at UIN Kediri. Its implementation can support continuous improvement in Arabic language testing services from technical, administrative, and participant-experience perspectives.

Discussion

Overall, this study produced an online TOAFL platform that met the criteria of feasibility, practicality, and effectiveness. The findings indicate that an LMS provides an appropriate medium for Arabic language proficiency testing because it offers flexibility, technical stability, and efficient administration. Compared with conventional test administration, the platform reduced the Language Center's technical burden, prevented participant congestion, and accelerated score processing.

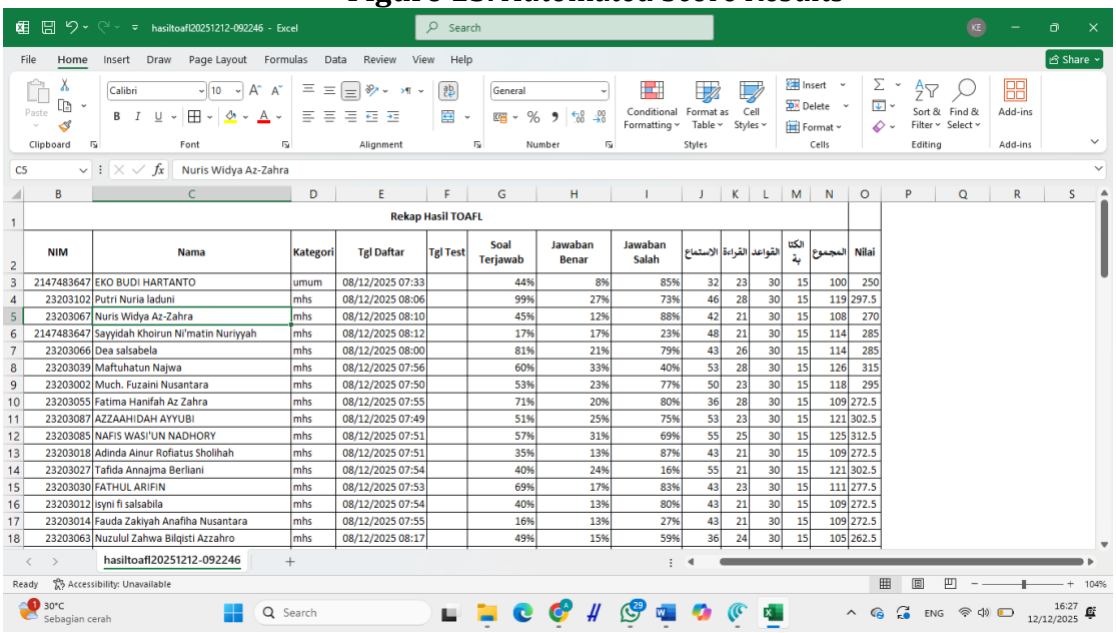
Figure 14. Automated TOAFL Score Summary



No	NIM	Nama	Kategori	Tgl Daftar	Tgl Test	Soal Terjawab	Jawaban Benar	Jawaban Salah	الاستماع	القراءة	القواعد	الكتابة	المجموع	Nilai
1	2147483647	EKO BUDI HARTANTO	umum	2025-12-08 07:33:18	2025-12-08 08:27:33	44 %	8 %	85 %	32	23	30	15	100	250
2	23203102	Putri Nuria laduni	mhs	2025-12-08 08:08:20	2025-12-08 08:29:04	99 %	27 %	73 %	46	28	30	15	119	297.5
3	23203067	Nuris Widya Az-Zahra	mhs	2025-12-08 08:10:39	2025-12-08 08:29:32	45 %	12 %	88 %	42	21	30	15	108	270
4	2147483647	Sayyidah Khoirun Ni'matin Nuriyyah	mhs	2025-12-08 08:12:10	2025-12-08 08:32:40	17 %	17 %	23 %	48	21	30	15	114	285
5	23203066	Dea salsabela	mhs	2025-12-08 08:00:30	2025-12-08 08:32:43	81 %	21 %	79 %	43	28	30	15	114	285
6	23203039	Mafruhatun Najwa	mhs	2025-12-08 07:56:18	2025-12-08 08:34:34	60 %	33 %	40 %	53	28	30	15	126	315
7	23203002	Much. Fuzaini Nusantara	mhs	2025-12-08 07:50:46	2025-12-08 08:34:45	53 %	23 %	77 %	50	23	30	15	118	295
8	23203055	Fatima Hanifah Az Zahra	mhs	2025-12-08 07:55:53	2025-12-08 08:34:52	71 %	20 %	80 %	38	28	30	15	109	272.5
9	23203087	AZZAAHIDAH AYYUBI	mhs	2025-12-08 07:49:16	2025-12-08 08:35:24	51 %	25 %	75 %	53	23	30	15	121	302.5

Furthermore, the structured and readily updateable item bank supports continuous improvement in test quality. Expert validation and participant responses provide a strong basis for permanent platform implementation and support its potential adoption by other higher education institutions under the Ministry of Religious Affairs.

Figure 15. Automated Score Results



Rekap Hasil TOAFL

NIM	Nama	Kategori	Tgl Daftar	Tgl Test	Soal Terjawab	Jawaban Benar	Jawaban Salah	الاستماع	القراءة	القواعد	الكتابة	المجموع	Nilai
2147483647	EKO BUDI HARTANTO	umum	08/12/2025 07:33		44%	8%	85%	32	23	30	15	100	250
23203102	Putri Nuria laduni	mhs	08/12/2025 08:06		99%	27%	73%	46	28	30	15	119	297.5
23203067	Nuris Widya Az-Zahra	mhs	08/12/2025 08:10		45%	12%	88%	42	21	30	15	108	270
2147483647	Sayyidah Khoirun Ni'matin Nuriyyah	mhs	08/12/2025 08:12		17%	17%	23%	48	21	30	15	114	285
23203066	Dea salsabela	mhs	08/12/2025 08:00		81%	21%	79%	43	26	30	15	114	285
23203039	Maftuhatun Najwa	mhs	08/12/2025 07:56		60%	33%	40%	53	28	30	15	126	315
23203002	Much. Fuzaini Nusantara	mhs	08/12/2025 07:50		53%	23%	77%	50	23	30	15	118	295
23203055	Fatima Hanifah Az Zahra	mhs	08/12/2025 07:55		71%	20%	80%	36	28	30	15	109	272.5
23203087	AZZAAHIDAH AYYUBI	mhs	08/12/2025 07:49		51%	25%	75%	53	23	30	15	121	302.5
23203085	NAFIS WASI'UN NADHORY	mhs	08/12/2025 07:51		57%	31%	69%	55	25	30	15	125	312.5
23203018	Adinda Ainur Rofiatus Sholihah	mhs	08/12/2025 07:51		35%	13%	87%	43	21	30	15	109	272.5
23203027	Tafida Annajma Berliani	mhs	08/12/2025 07:54		40%	24%	16%	55	21	30	15	121	302.5
23203030	FATHUL ARIFIN	mhs	08/12/2025 07:53		69%	17%	83%	43	23	30	15	111	277.5
23203012	isyni fi salsabila	mhs	08/12/2025 07:54		40%	13%	80%	43	21	30	15	109	272.5
23203014	Fauda Zakiyah Anafiha Nusantara	mhs	08/12/2025 07:55		16%	13%	27%	43	21	30	15	109	272.5
23203063	Nuzulul Zahwa Bilqisti Azzahro	mhs	08/12/2025 08:17		49%	15%	59%	36	24	30	15	105	262.5

The implementation of the LMS-based TOAFL platform represents a significant paradigm shift from conventional, centralized Computer-Based Testing (CBT) to a decentralized *Test from Anywhere* (TfA) model. As highlighted by Ockey (2009), traditional CBT systems are frequently bottlenecked by the physical constraints of institutional infrastructure. This study demonstrates that utilizing an LMS circumvents the severe laboratory capacity limitations at UIN Syekh Wasil Kediri, allowing the simultaneous assessment of massive examinee cohorts without physical congestion or repeated scheduling. By reducing the administrative workload and eliminating spatial dependencies, this innovation aligns with Sarwadi et al. (2025), confirming that LMS environments fundamentally enhance the operational scalability, flexibility, and accessibility of language proficiency assessments.

A critical concern in remote testing is maintaining academic integrity and assessment validity. This study addressed this challenge by embedding strict security protocols within the LMS, including randomized item sequencing, shuffled response options, strict time limits, and restricted backward navigation when necessary. These features collectively minimize the opportunities for academic dishonesty that often threaten unproctored online assessments. Furthermore, in line with Hamp-Lyons' (2017) and Brown's (2006) assertions regarding fairness and objectivity, the platform's intuitive user interface and clear navigation serve to reduce construct-irrelevant variance. By ensuring that examinees are solely tested on their Arabic proficiency rather than their ability to navigate a convoluted digital system, the platform safeguards the construct validity of the TOAFL scores.

Beyond technical delivery, the robustness of the TOAFL system is rooted in the systematic development and quality of its integrated item bank. The item analysis revealed a proportional distribution of difficulty, with most items falling into the moderate category, coupled with strong discrimination indices. As Alderson (2010)

posits, this concentration of moderately difficult items is essential for proficiency tests, as it maximizes the instrument's capacity to differentiate between varying levels of language competence accurately. The strategic decision to iteratively revise—rather than immediately discard—items with extreme difficulty or poor discrimination underscores a sustainable approach to test development. This continuous refinement cycle ensures the longevity and dynamic expansion of the item bank across successive testing periods, maintaining test equivalence and preventing item exposure.

Ultimately, the successful deployment of this platform extends its implications far beyond administrative convenience, offering profound institutional and pedagogical value. The automated scoring and detailed participant activity logs provide faculty and administrators with granular data that can inform data-driven decisions regarding Arabic language curricula. Furthermore, the standardized nature of this LMS-based TOAFL presents a highly scalable model that can be adopted by other higher education institutions under the Ministry of Religious Affairs. Looking forward, the establishment of this validated, digitized item bank lays the crucial empirical groundwork for a future transition toward Computerized Adaptive Testing (CAT), which would further optimize testing efficiency, security, and precision in measuring Arabic language proficiency on a global scale.

CONCLUSION

This study produced a Learning Management System based Test of Arabic as a Foreign Language (TOAFL) platform that addressed the institutional need for a more flexible, efficient, and accessible Arabic language proficiency assessment system. Unlike conventional Computer-Based Testing, the developed platform reduced dependence on computer laboratories, supported the management of large participant groups, and provided activity logging, time configuration, and automated scoring. The study also developed the TOAFL item bank systematically through blueprint preparation, test specification, item writing, and item quality analysis. The item bank covered four competence components, namely istimā', qirā'ah, qawā'id, and insyā', and therefore represented Arabic proficiency comprehensively.

Expert validation, product trials, and item quality analysis indicated that the TOAFL platform and item bank ranged from feasible to highly feasible for implementation. The platform remained stable, was easy to use, and supported simultaneous access from different devices. Most items demonstrated moderate difficulty and good to very good discrimination. The study therefore produced a valid, practical, and effective LMS-based TOAFL assessment system with the potential to serve as a model for standardized and sustainable Arabic language assessment in Islamic higher education institutions.

This study offers practical implications for the development of Arabic language proficiency assessment in higher education. In particular, the LMS-based TOAFL platform provides a more flexible, efficient, and accessible alternative to laboratory-dependent testing. The platform can reduce reliance on institutional computer laboratories, improve test-management efficiency, and support a more objective and documented scoring process. The systematically developed item bank can also provide a foundation for a standardized and sustainable TOAFL instrument, thereby improving the quality of Arabic language assessment services in Islamic higher education institutions.

Nevertheless, this study has several limitations. The researchers developed and tested the product only at UIN Syekh Wasil Kediri, which limits the generalizability of the findings to other institutions. The empirical validation of the instrument focused on difficulty and discrimination analyses based on classical test theory and did not include more comprehensive psychometric evaluation, such as reliability analysis, construct validity, or Item Response Theory modeling. Moreover, the platform evaluation involved only short-term trials and therefore did not fully demonstrate its long-term effectiveness in regular TOAFL administration.

Future studies should implement the platform across several higher education institutions with more diverse user characteristics to examine its external validity. Researchers should also expand the item bank and conduct more comprehensive psychometric analyses, including reliability testing, construct validation, and item calibration through Item Response Theory. Further evaluation of user experience, system security, platform adaptivity, and LMS-based computerized adaptive testing represents a promising research agenda for improving the accuracy, efficiency, and sustainability of future TOAFL administration.

REFERENCES

Abdulquadir, Ahmad Ayobami, Opeyemi Abdulsalam, Russell Mcmahon, and Annu Prabhakar. 2023. "Usability Study of a Learning Management System (LMS)." *The 24th Annual Conference on Information Technology Education*, October 11, 150–56. https://doi.org/10.1145/3585059.3611415.

Afyuddin, Moh Sholeh, Maziyyatul Muslimah, Muhammad Nur Kholis, and Rina Dian Rahmawati. 2023. "Development of Tadris Al-'Arabiyyah Al-Mukatssaf (TAM) Textbook, Based on Contextual Teaching and Learning (CTL)." *Asalibuna* 7 (01): 1–17. https://doi.org/10.30762/asalibuna.v7i01.1084.

Ahmad, Muhammad Abdul Qodir. 1979. *Thuruqu Ta'lim al Lughah al Arobiyah*. Maktabah Annahdhoh.

Al Muchith, Aim Matun Nihayati and Nely Rahmawati Zaimah. 2023. "A Relevance of Learning a Foreign Language for an Early Age." *Mantiqu Tayr: Journal of Arabic Language* 3 (1): 25–35. https://doi.org/10.25217/mantiqutayr.v3i1.3117.

Albantani, Azkia Muharom, and Imam Fitri Rahmadi. 2020. "Mobile Devices for Arabic Learning in Junior High Schools: The Teachers' Perspective." *Jurnal Al Bayan: Jurnal Jurusan Pendidikan Bahasa Arab* 12 (2): 191–207. https://doi.org/10.24042/albayan.v12i2.6385.

Alderson, J. Charles. 2000. *Assessing Reading*. 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935.

Alderson, John Charles, ed. 2010a. *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. Continuum.

Alderson, John Charles, ed. 2010b. *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. Continuum.

Al-Fraihat, Dimah, Mike Joy, Ra'ed Masa'deh, and Jane Sinclair. 2020. "Evaluating E-Learning Systems Success: An Empirical Study." *Computers in Human Behavior* 102 (January): 67–86. https://doi.org/10.1016/j.chb.2019.08.004.

Alhawary, Mohammad T. 2019. *Arabic Second Language Learning and Effects of Input, Transfer, and Typology*. Georgetown University Press. https://doi.org/10.1353/book63892.

Anwar, Saiful, Guntur Cahaya Kesuma, and Koderi. 2023. "Development of Al-Qawaid an-Nahwiyah Learning Module Based on Qiyasiyah Method for Arabic Language Education Department Students | Pengembangan Modul Pembelajaran al-Qawaid an-Nahwiyah Berbasis Metode Qiyasiyah Untuk Mahasiswa Jurusan Pendidikan Bahasa Arab." *Mantiqu Tayr: Journal of Arabic Language* 3 (1): 11–24. https://doi.org/10.25217/mantiqutayr.v3i1.2830.

Black, James. n.d. *Metode Dan Masalah Penelitian Sosial*. Refika Aditama.

Branch, Robert Maribe. 2009. *Instructional Design: The ADDIE Approach*. Springer US. https://doi.org/10.1007/978-0-387-09506-6.

Brown, H. Douglas. 2006a. *Language Assessment: Principles and Classroom Practices*. Longman.

Brown, H. Douglas. 2006b. *Language Assessment: Principles and Classroom Practices*. Longman.

Buck, Gary. 2001. *Assessing Listening*. The Cambridge Language Assessment Series. Cambridge University Press. https://doi.org/10.1017/CBO9780511732959.

Cao, Thi Xuan Lien. 2023. "Benefits and Challenges of Using LMS in Blended Learning: Views from EFL Teachers and Students at a Vietnamese Public University." *International Journal of TESOL & Education* 3 (3): 78–100. https://doi.org/10.54855/ijte.23335.

Hamid, M. Abdul, and M. Abdullah Charis. 2021. "Development Of Computer Based Test Online Standard Arabic Language (ILAA)/ Pengembangan Aplikasi Computer Based Test Online Bahasa Arab Terstandar (ILAA)." *Ijaz Arabi Journal of Arabic Learning* 4 (2). https://doi.org/10.18860/ijazarabi.v4i2.11194.

Hamp-Lyons, Liz. 2017. "Language Assessment Literacy for Language Learning-Oriented Assessment." *Studies in Language Assessment*, 88–111. https://doi.org/10.58379/LIXL1198.

Hasan, Hasan. 2018. "KETERAMPILAN MENGAJAR BAHASA ARAB MATERI ISTIMA MENGGUNAKAN MEDIA LAGU." *Al Qalam: Jurnal Ilmiah Keagamaan Dan Kemasyarakatan*, October 5, 127. https://doi.org/10.35931/aq.v0i0.7.

Herniawati, Ani, Laskmi Holifah, and Abd. Syakur. 2025. "Effectiveness of Learning Management System (LMS) Use in Higher Education." *International Journal Corner of Educational Research* 4 (1): 20–29. https://doi.org/10.54012/ijcer.v4i1.625.

Kane, Michael T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50 (1): 1–73. https://doi.org/10.1111/jedm.12000.

Kasabova, Gergana, Silvia Parusheva, University of Economics - Varna, Varna, Bulgaria, Boris Bankov, and University of Economics - Varna, Varna, Bulgaria. 2023. "Learning Management Systems as a Tool for Learning in Higher Education." *Izvestia Journal of the Union of Scientists - Varna. Economic Sciences Series* 12 (2): 224–33. https://doi.org/10.56065/IJUSV-ESS/2023.12.2.224.

Koentjaraningrat. 1994. *Metode-Metode Penelitian Masyarakat*. Gramedia.

Margono, S. 1997. *Metodologi Penelitian Pendidikan*. PT. Rineka Cipta.

Maul, Asmaul Husna, and Inchinia Angger Rowin. 2026. "Developing Digital Learning Tools Based on 21st-Century Skills for the Arabic Speaking Skill at Maulana Malik Ibrahim State Islamic University Malang." *Asalibuna* 9 (02): 194–211. https://doi.org/10.30762/asalibuna.v9i02.7007.

Moh. Ismail; Aufa Alfian Musthafa. 2018. "Pengembangan Materi Tes Keterampilan Menyimak Bahasa Arab Menggunakan Moodle Untuk Meningkatkan Kemampuan Menyimak Mahasiswa Universitas Darussalam Gontor." *At-Ta'dib*, no. Vol 13, No 2 (2018): Pesantren as a Center for Developing Arabic&English Teaching: 28–49.

Moleong, Lexy J. 2001. *Metode Penelitian Kualitatif*. PT. Remaja Rosdakarya.

N. Hula, Ibnu Rawandhy, Moh. Zulkifli Paputungan, and Ana Mariana. 2021. "PENGEMBANGAN HYBRID LEARNING BERBASIS APLIKASI COMPUTER ASSITED TEST (CAT) PADA PROGRAM ARABIC PROFICIENCY TEST." *Tadbir: Jurnal Manajemen Pendidikan Islam* 9 (1): 103–25. https://doi.org/10.30603/tjmpi.v9i1.2063.

Naqoh, Mahmud Kamil al-. 2006. *Ta'lim al-Lughoh Ittishaliyan Baina al-Manahij Wa al-Istiratijiyat*. Issesco.

Nation, I. S. P. 2006. *Learning Vocabulary in Another Language*. 8. print. The Cambridge Applied Linguistics Series. Cambridge Univ. Press.

Nurdianto, Talqis, Yayat Hidayat, and Ibnu Ali Abdullah Fitrianto. 2025. "Design of CEFR-Based Arabic Language Proficiency Test Indicators In The Arabic Language Education Department." *Ijaz Arabi Journal of Arabic Learning* 8 (3). https://doi.org/10.18860/ijazarabi.v8i3.28795.

Ockey, Gary J. 2009. "Developments and Challenges in the Use of Computer-Based Testing for Assessing Second Language Ability." *The Modern Language Journal* 93 (s1): 836–47. https://doi.org/10.1111/j.1540-4781.2009.00976.x.

Qodri, Muhammad. 2020. "Problematika Pembelajaran TOAFL Pada Mahasiswa Fakultas Ilmu Tarbiyah Dan Keguruan UIN Sulthan Thaha Saifuddin Jambi." *Loghat Arabi : Jurnal Bahasa Arab Dan Pendidikan Bahasa Arab* 1 (1): 1. https://doi.org/10.36915/la.v1i1.1.

Rosmiati, Ucu, Supratman Supratman, and Sri Tirto Madawistama. 2021. "ELMA (E-Learning for Matheatics): Pengembangan Media Pembelajaran Berbasis LMS (Learning Management System) Sebagai Alternatif Pembelajaran Di Masa Pandemi Covid-19." *JIPM (Jurnal Ilmiah Pendidikan Matematika)* 10 (2): 163. https://doi.org/10.25273/jipm.v10i2.10051.

Sa'idah, Ratna, Patria Ayudha Putri, Fatma Fitriana, Umar Faruq, and Eko Budi Hartanto. 2022. "Intensive Arabic Language Program to Strengthen the Reading Ability of New Arabic Language Education Students." *Asalibuna* 6 (2): 1–9.

Sarwadi, Sarwadi, Sayaiful Huda, Mujtahidin Mujtahidin, and Bohri Rahman. 2025. "The Impact of Learning Management Systems (LMS) on University Students' Language Skills." *Jurnal Pendidikan Bahasa Inggris Undiksha* 13 (1): 116–23. https://doi.org/10.23887/jpbi.v13i1.93390.

Sommerville, Ian. 2016. *Software Engineering*. 10. ed. Pearson.

Syalabi, Ahmad. 1997. *Kaifa Taktub Baḥsan Aw Risâlah?* Maktabah al-Nahḍah al-Miṣriyah.

Tahir, Mohd Haniff Mohd, Syahwil Saputra, Salwa Othman, et al. 2025. "Online Assessment in Higher Education: A Systematic Literature Review." *Multidisciplinary Reviews* 9 (1): 2026024. https://doi.org/10.31893/multirev.2026024.

Thiagarajan, S., Dorothy S. Semmel, and Melvyn I. Semmel. 1974. *Instructional Development for Training Teachers of Exceptional Children*. Indiana University.

Ulya, Nia Himatul, Chairani Astina, and Ashief El Qorny. 2022. "Implementation of Bi'ah Lughawiyah in Improving Maharah Kalam at Modern Pondok Az-Zahra al-Gontory Purwokerto|Implementasi Bi'ah Lughawiyyah Dalam Peningkatan Maharah Kalam Di Pondok Modern Az-Zahra al-Gontory Purwokerto." *Mantiqu Tayr: Journal of Arabic Language* 2 (2): 174–84. https://doi.org/10.25217/mantiqutayr.v2i2.2511.

Watson, William R. n.d. "An Argument for Clarity: What Are Learning Management Systems, What Are They Not, and What Should They Become." *TechTrends* 51 (2): 28–34.

Yusuf Qathith, Ghassan. 2009. *Al-Hasub Wa Turuqu At-Tadris Wa At-Taqwim*. Dar al-Tsaqafah Wa Nashr.

Zubaidi, Ahmad. 2020. "DEVELOPMENT OF MAHÂRAH AL-ISTIMÂ' TEST INSTRUMENT FOR ELECTRONIC BASED ARABIC STUDENT USING THE KAHOOT! APPLICATION." *Arabiyat : Jurnal Pendidikan Bahasa Arab Dan Kebahasaaraban* 7 (1): 61–77. https://doi.org/10.15408/a.v7i1.13002.