DALWA REVIEW OF LANGUAGE AND LITERACY
DRLL | Vol. 1 No. 1 | 2026 | pp. 56 - 70
ISSN: XXXX-XXXX (Online) | DOI: 10.XXXX/XXXXXXX

Evaluating the Test of Arabic as a Foreign Language (TOAFL) Graduation Requirement Policy in Higher Education: Students Perspectives

Muhammad Arif Nasruddin[1], **Afif Kholisun Nashoih**[2]
[1] *Universitas Islam Raden Rahmat, Malang, Indonesia*
[2] *Universitas KH. A. Wahab Hasbullah, Jombang, Indonesia*
✉ **Corresponding author:** afifkholis@unwaha.ac.id

Received:	30 March 2026
Accepted:	9 May 2026
Published:	9 May 2026
DOI:	https://doi.org/10.XXXX/drll.XXXX

ABSTRACT

The Test of Arabic as a Foreign Language (TOAFL) graduation requirement policy in Indonesian Islamic universities has attracted increasing academic attention. This study evaluates the implementation of the TOAFL policy from students' perspectives, encompassing their perceptions of policy relevance, barriers encountered during preparation, and the policy's impact on Arabic language learning motivation. A mixed-methods design was employed, combining a validated 48-item questionnaire administered to 312 active students at five state Islamic universities across Indonesia with semi-structured interviews conducted with 24 purposively selected participants. Quantitative data were analysed using multiple linear regression, MANOVA, and one-way ANOVA with Bonferroni post-hoc correction; qualitative data were subjected to reflexive thematic analysis. Findings indicate that the majority of students (67.3%) acknowledge TOAFL's relevance for linguistic competence development, yet a significant gap exists between the quality of classroom Arabic instruction and the competence standards demanded by the examination. Access to learning resources, instructor qualifications, and examination frequency significantly predicted students' TOAFL attainment. Psychometric reviews of the TOAFL instrument within the CEFR framework (Anggara, 2026) further indicate that the instrument requires systematic refinement to align with internationally recognised language proficiency standards. The study recommends policy reforms encompassing strengthened instructional infrastructure, standardisation of Arabic instruction, and contextually sensitive revision of passing thresholds.

Keywords: TOAFL; Policy Evaluation; Arabic Language; Islamic Higher Education; Students' Perspectives.

INTRODUCTION

Arabic language proficiency constitutes a strategic competency expected of graduates from Islamic universities, given its central role as the language of the Qur'an, the Hadith, and the rich corpus of classical Islamic scholarship. To ensure that this linguistic standard is met, a growing number of Islamic higher education institutions in Indonesia-including State Islamic Universities (Universitas Islam Negeri/UIN), State Institutes for Islamic Studies (Institut Agama Islam Negeri/IAIN), and Islamic Studies Colleges (Sekolah Tinggi Agama Islam/STAI)-have adopted the policy of requiring students to pass the Test of Arabic as a Foreign Language (TOAFL) as a prerequisite for graduation. While grounded in legitimate academic objectives, this policy has generated considerable debate regarding the fairness and feasibility of its implementation, particularly from the perspective of students as its most directly affected stakeholders (Shohamy, 2001; Rosyidi, 2021).

The evaluation of language-in-education policies-including the use of standardised tests as graduation gatekeepers-has received extensive attention in the applied linguistics literature (Alderson, 2004; Shohamy, 2001). The washback effect of standardised testing on learning processes, student motivation, and classroom instructional practices represents one of the central concerns of language assessment policy research (Hughes, 2003; Wall, 2005). In Southeast Asian contexts, research on analogous policies-such as compulsory TOEFL or IELTS requirements at certain universities-demonstrates that although such requirements may promote language improvement, their implementation frequently generates inequitable outcomes and undue psychological pressure for specific student subgroups (Hamid & Baldauf, 2013; Kirkpatrick, 2011). Psychometric reviews of the TOAFL instrument itself indicate that it requires systematic alignment with the Common European Framework of Reference for Languages (CEFR) to measure Arabic proficiency in a more representative and equitable manner (Anggara, 2026).

Dedicated empirical studies on TOAFL in Indonesia remain limited. Existing research has focused predominantly on the psychometric properties of the instrument (Mahyudin, 2019; Anggara, 2026) or single-institution implementation descriptions (Rosyidi, 2021), thereby failing to provide a comprehensive evaluative picture from students' perspectives across multiple institutions. This gap forms the primary rationale for the present study. Three research questions guide the inquiry: (1) How do students perceive the relevance and fairness of the TOAFL policy? (2) What are the primary barriers students face in preparing for TOAFL? (3) How does the TOAFL requirement affect students' motivation to learn Arabic?

Answers to these questions are relevant not only to policymakers at Islamic higher education institutions but also carry broader implications for the design of Arabic language assessment instruments that are

valid, fair, and predictively powerful. By integrating quantitative and qualitative approaches, this study seeks to generate evidence-based recommendations to assist institutions in crafting TOAFL policies that are more effective, equitable, and responsive to students' diverse needs.

LITERATURE REVIEW

2.1. Standardised Testing as Language Policy

The use of standardised tests as policy instruments in higher education has long been debated in the applied linguistics literature. Shohamy (2001) argues that standardised tests carry ideological power that far exceeds their measurement function; they operate as instruments of social control capable of shaping curriculum, instructional practices, and student learning behaviours in their entirety. This phenomenon is termed the 'washback' or 'backwash effect'-the indirect influence of testing policy on the processes of teaching and learning (Hughes, 2003; Bailey, 1996).

In foreign language contexts, positive washback manifests when examinations encourage students to develop holistic linguistic competence. Negative washback, conversely, occurs when students and instructors focus exclusively on test preparation (test-oriented learning) at the expense of authentic communicative development (Cheng, 2005). McNamara and Roever (2006) underscore the critical importance of equitable access (fairness) in standardised testing, particularly for student populations whose prior Arabic language education has been uneven. This concern becomes especially pressing when standardised tests serve as high-stakes graduation gates rather than merely diagnostic instruments-as is the case with TOAFL in Indonesian Islamic universities.

2.2. TOAFL: Instrument, Psychometrics, and the CEFR Framework

TOAFL was originally developed by the Language Development Centre of UIN Syarif Hidayatullah Jakarta in the early 2000s as a contextualised measure of Arabic language ability adapted to the academic needs of Indonesian university students (Zulhannan, 2014). Structurally, TOAFL assesses four core components: (1) listening comprehension, (2) written structure and expression, (3) reading comprehension, and (4) contextual vocabulary. Score scales range from 217 to 677, with passing thresholds varying across institutions, typically between 400 and 450.

A systematic literature review conducted by Anggara (2026) examining psychometric studies of the TOAFL instrument within the CEFR framework revealed several important findings. First, TOAFL generally demonstrates adequate internal reliability (Cronbach's alpha > .80), though the construct validity of inter-subtest relationships requires further investigation. Second, mapping the difficulty levels of TOAFL items onto CEFR descriptors indicates that the majority of items fall within the A2–B1 range, whereas institutional proficiency targets typically aim for B2-level competence. Third, a representational gap exists in productive skills-particularly speaking and writing-which are not assessed in the standard TOAFL

format. These findings indicate the need for fundamental reform in TOAFL instrument design to align with internationally recognised Arabic language proficiency standards (Anggara, 2026).

The adoption of TOAFL as a graduation requirement has expanded among Islamic higher education institutions since the mid-2010s, in parallel with the internationalisation agenda of Islamic higher education and efforts to enhance graduate quality (Hidayat, 2018). Nonetheless, systematic evaluation of the policy's impact-particularly from the perspective of students as primary users-remains sparse in the academic literature in both Indonesian and English.

2.3. Arabic Learning Motivation: Theoretical Perspectives

Motivation in foreign language learning can be understood through the Self-Determination Theory (SDT) framework developed by Deci and Ryan (2000). SDT distinguishes between intrinsic motivation-arising from the inherent pleasure and satisfaction of the learning activity itself-and extrinsic motivation, driven by external factors such as grades, recognition, or, in the present context, the obligation to pass an examination. Research by Noels et al. (2000) demonstrates that autonomy-supportive learning environments enhance intrinsic motivation, while excessive external pressure tends to erode self-confidence and long-term learning engagement.

In the context of Arabic as a religious language, instrumental and integrative motivational dynamics are particularly distinctive. Students in Islamic higher education contexts tend to harbour strong integrative motivation-the desire to understand religious texts directly-yet frequently encounter structural barriers in the form of limited exposure to Arabic outside formal classroom instruction (Wahyudi, 2020). The TOAFL policy, if implemented without adequate learning support, risks shifting students' motivational orientation from integrative to purely instrumental, with potentially negative consequences for the depth and sustainability of Arabic language acquisition.

Dörnyei's (2009) Ideal L2 Self construct offers an additional relevant lens: students who maintain a vivid imagined future self as a competent Arabic speaker will invest greater effort in learning, including in TOAFL preparation. This implies that institutions must do more than set passing standards; they must actively cultivate an affective ecosystem that connects Arabic language mastery to students' professional and spiritual aspirations.

2.4. Theoretical Framework

This study is grounded in an integrative theoretical framework combining three perspectives. First, Critical Language Testing (CLT) from Shohamy (2001), which emphasises the importance of analysing assessment policies through the lens of power, fairness, and social impact. Second, Self-Determination Theory (SDT) from Deci and Ryan (2000), which provides a framework for understanding student motivational dynamics in the context of externally imposed evaluative policies. Third, the Common

European Framework of Reference for Languages (CEFR), which serves as an international standard against which the adequacy of the TOAFL instrument can be evaluated-as systematically developed in the review by Anggara (2026).

The integration of CLT and SDT yields a distinctive perspective: the TOAFL policy is understood not solely as a measurement mechanism, but also as a motivational intervention capable of either building or undermining students' learning orientations depending on how the policy is designed, communicated, and supported by the institutional learning ecosystem. The CEFR provides the third vertex of the framework-a normative benchmark for assessing whether the competence measured by TOAFL reflects internationally recognised proficiency and whether the graduation thresholds set are realistic and assessable.

This tripartite conceptual framework can be visualised as a triangle: (1) the TOAFL policy (including the instrument and passing standards) as the structural factor; (2) student perceptions, barriers, and motivations as individual factors; and (3) the institutional Arabic language learning ecosystem as the mediating factor. All three vertices are interdependent: a strong policy without a supportive ecosystem generates barriers; a sound ecosystem without a coherent policy loses direction; and positive individual factors are optimised only when both structural and ecological conditions are adequately met.

METHOD

3.1. Research Design and Participants

This study employed a mixed-methods research design following an explanatory sequential strategy, in which quantitative data were collected and analysed first, followed by qualitative data collection to explore and deepen quantitative findings (Creswell & Plano Clark, 2018). This approach enables triangulation between statistical data and the lived narrative experiences of students, yielding a more comprehensive account of the TOAFL policy's impact. Stratified random sampling was applied based on institution type and cohort year.

Participants comprised 312 active students drawn from five state Islamic universities across Indonesia: UIN Maulana Malik Ibrahim Malang, UIN Sunan Kalijaga Yogyakarta, IAIN Ponorogo, IAIN Palangka Raya, and IAIN Syekh Nurjati Cirebon. The gender distribution was 54.2% female and 45.8% male, with an age range of 18–26 years (M = 21.3; SD = 1.74). Of the sample, 63.5% were enrolled in programmes requiring TOAFL as a prerequisite for the comprehensive examination, while 36.5% came from programmes where TOAFL was required as a graduation prerequisite. Data were collected online between February and April 2024. Detailed participant characteristics are presented in Table 1.

Table 1. *Participant Characteristics (N = 312)*

Characteristic	Frequency (n)	Percentage (%)
Gender: Female	169	54.2
Gender: Male	143	45.8
Age: 18–20 years	134	43.0
Age: 21–23 years	148	47.4
Age: 24–26 years	30	9.6
Programme: TOAFL compulsory	198	63.5
Programme: TOAFL as graduation prerequisite	114	36.5
Background: Islamic boarding school alumni	127	40.7
Background: Non-boarding school	185	59.3

Note. *Islamic boarding school (pesantren) classification includes students who resided at a pesantren for at least two years.*

3.2. Instruments

The primary research instrument was a researcher-developed questionnaire validated through expert judgement by three specialists (applied linguistics, educational measurement, and Islamic education policy). The questionnaire comprised 48 items across four subscales: (1) Perception of Policy Relevance (12 items; $\alpha = .84$), (2) Barriers to TOAFL Preparation (14 items; $\alpha = .88$), (3) Extrinsic Motivational Impact (11 items; $\alpha = .81$), and (4) Intrinsic Motivational Impact (11 items; $\alpha = .79$). All items employed a 4-point Likert scale (1 = Strongly Disagree to 4 = Strongly Agree). Content validity was assessed using Lawshe's (1975) Content Validity Ratio (CVR) across the three expert raters, yielding a mean CVR of .87, which exceeds the critical value for three raters.

The TOAFL instrument evaluated in this study references the standard format developed by the Language Development Centre of UIN Maliki Malang, which has been subjected to psychometric review within the CEFR framework (Anggara, 2026). That review provides an essential empirical backdrop for understanding the characteristics of the instrument being evaluated from students' perspectives. For the qualitative phase, semi-structured interviews (40–55 minutes each) were conducted with 24 purposively selected students stratified by questionnaire score profile, institutional affiliation, and prior TOAFL examination experience.

The interview protocol was structured around three thematic domains: (a) personal narrative of TOAFL preparation experience and strategies; (b) perceived institutional support and barriers; and (c) the perceived relationship between TOAFL requirements and broader Arabic language learning motivation. All interviews were audio-recorded with participants' informed consent, transcribed verbatim, and member-checked with six participants to verify interpretive accuracy. Ethical clearance was obtained from the institutional review boards of both authors' home institutions prior to data collection, and all

participants provided written informed consent. Participant confidentiality was maintained through the use of alphanumeric identifiers throughout data reporting.

3.3. Data Analysis

Quantitative data were analysed using IBM SPSS Statistics v.28. Multiple linear regression addressed RQ1, with policy relevance perception as the predictor and examination preparedness as the dependent variable. MANOVA examined the effect of learning resource accessibility on TOAFL attainment (RQ2). One-way ANOVA with Bonferroni post-hoc correction compared motivational impact across student groups (RQ3). Normality (Shapiro-Wilk) and homogeneity of variance (Levene) assumptions were verified prior to all inferential tests. Qualitative data from interview transcripts were analysed using reflexive thematic analysis (Braun & Clarke, 2006) with NVivo 14, achieving satisfactory intercoder agreement (Cohen's $\kappa = .81$) on a 25% subsample coded independently by both researchers.

FINDINGS

4.1. Descriptive Statistics

Descriptive analysis revealed that the Barriers to TOAFL Preparation subscale recorded the highest mean among the four subscales ($M = 2.95$; $SD = 0.71$), indicating that students face substantial challenges in examination preparation. The Perception of Policy Relevance subscale obtained a mean of $M = 2.81$ ($SD = 0.67$), while Extrinsic Motivational Impact ($M = 2.68$; $SD = 0.74$) and Intrinsic Motivational Impact ($M = 2.23$; $SD = 0.82$) showed more varied patterns. Table 2 presents the complete descriptive statistics.

Table 2. *Descriptive Statistics for Questionnaire Subscales (N = 312)*

Variable / Subscale	Min.	Max.	M	SD
Perception of Policy Relevance	1.00	4.00	2.81	0.67
Barriers to TOAFL Preparation	1.00	4.00	2.95	0.71
Extrinsic Motivational Impact	1.00	4.00	2.68	0.74
Intrinsic Motivational Impact	1.00	3.75	2.23	0.82

Note. *Response scale: 1 = Strongly Disagree, 4 = Strongly Agree.*

4.2. Findings for RQ1: Perception of Policy Relevance

Multiple regression analysis demonstrated that perception of policy relevance significantly predicted examination preparedness ($\beta = .42$, $t = 7.83$, $p < .001$, $R^2 = .31$). Prior Arabic language educational background (pesantren vs. non-pesantren) also contributed a significant predictive effect ($\beta = .28$, $t = 5.17$, $p < .001$), providing support for H1. The overall regression model was significant, $F(4, 307) = 34.72$, $p < .001$, $R^2 = .31$, accounting for 31% of variance in examination preparedness.

Further analysis revealed that pesantren alumni scored an average of 8.4 points higher in examination preparedness compared to non-pesantren students (95% CI [6.1, 10.7], $p < .001$). This confirms that a

uniform TOAFL passing threshold risks creating structural inequity for students who lack access to intensive pre-university Arabic language education-a finding convergent with the psychometric concerns raised by Anggara (2026) regarding the need for instrument adaptation to accommodate diverse proficiency profiles.

4.3. Findings for RQ2: Barriers to TOAFL Preparation

MANOVA revealed a significant main effect of learning resource accessibility on TOAFL attainment, $F(8, 608) = 9.47$, $p < .001$, $\eta^2 = .11$ (medium effect size). Subsequent univariate analyses identified Arabic instructional quality, $F(2, 309) = 22.14$, $p < .001$, and availability of relevant practice materials, $F(2, 309) = 17.89$, $p < .001$, as the strongest predictors. Table 3 summarises the full MANOVA results.

Table 3. *MANOVA Summary: Effect of Learning Resource Accessibility on TOAFL Attainment*

Independent Variable	F	df	p	η^2
Quality of Arabic Language Instruction	22.14	2, 309	< .001	.13
Availability of Practice Materials	17.89	2, 309	< .001	.10
Frequency of Examination Opportunities	11.23	2, 309	< .001	.07
Instructor Qualifications	9.87	2, 309	< .001	.06
Overall Model	9.47	8, 608	< .001	.11

Note. *Wilks' Lambda test. η^2 = partial eta-squared. Bonferroni correction applied for multiple comparisons.*

Thematic analysis of interview data identified four principal barrier themes: (a) insufficient time for independent Arabic study outside formal coursework, (b) limited access to standardised practice materials aligned with the TOAFL format, (c) perceived misalignment between classroom instruction content and TOAFL examination demands, and (d) psychological pressure associated with the consequences of repeated failure on academic progression. These themes were mutually reinforcing, creating a cycle of disadvantage difficult to interrupt without systemic intervention. Table 4 presents the qualitative themes alongside their frequency of occurrence across the interview sample.

Table 4. *Qualitative Themes: Barriers to TOAFL Preparation (n = 24 Interviewees)*

Theme	Sub-themes	Frequency	Representation (%)
Insufficient independent study time	Academic workload; part-time employment	21	87.5
Limited practice material access	Item format; Arabic academic vocabulary	19	79.2
Instruction–examination misalignment	Curriculum not aligned with TOAFL format	18	75.0
Psychological pressure	Test anxiety; fear of failure; procrastination	16	66.7
Unclear passing threshold standards	Policy changes across cohorts	11	45.8

Note. *Frequency counts reflect the number of participants who mentioned the theme during interview.*

4.4. Findings for RQ3: Motivational Impact

One-way ANOVA revealed significant between-group differences in extrinsic motivational impact by programme of study, $F(4, 307) = 8.93$, $p < .001$. Bonferroni post-hoc tests indicated that students in the Arabic Language Education programme recorded significantly higher extrinsic motivational impact scores than students in non-language programmes ($p < .05$), suggesting that personal relevance to Arabic moderates the intensity of the TOAFL policy's extrinsic washback.

Intrinsic motivational impact showed a more complex pattern: 41.7% of respondents reported increased interest in Arabic language learning after understanding the TOAFL requirements, while 38.5% reported experiences of language anxiety that interfered with their learning process-particularly among students who had failed the examination more than once. The remaining 19.8% reported no significant motivational impact in either direction, possibly reflecting psychological adaptation to a long-standing external evaluative policy.

Cross-analysis of quantitative and qualitative data revealed a noteworthy divergence: pesantren alumni with strong integrative motivation reported predominantly positive washback-the TOAFL policy reinforced rather than distorted their Arabic learning orientation. By contrast, non-pesantren students who encountered TOAFL solely as an institutional obligation were more susceptible to fragile extrinsic motivation-strong during examination pressure, but quick to diminish once the target was achieved. This divergence carries significant implications for how institutions communicate and contextualise the TOAFL policy to incoming students.

4.5. Integrated Predictive Model

To synthesise findings from the three research questions into a unified predictive account, a hierarchical regression analysis was conducted with TOAFL examination preparedness as the dependent variable. Table 5 presents the regression coefficients for the four key predictors identified through prior bivariate analyses.

Table 5. *Hierarchical Regression Analysis Summary: Predictors of TOAFL Examination Preparedness (N = 312)*

Predictor	β	SE	t	p	ΔR^2
Model 1: Pesantren background	.28	.04	5.17	< .001	.14
Model 2: + Perception of policy relevance	.42	.05	7.83	< .001	.17
Model 3: + Instructional quality	.31	.06	5.48	< .001	.08
Model 3: + Availability of practice materials	.24	.06	4.12	< .001	-
Model 3: + Extrinsic motivational impact	.19	.05	3.67	< .001	-
Final model ($R^2 = .39$; $F = 29.4$, $p < .001$)					.39

Note. *β = standardised beta coefficient; SE = standard error. All models controlled for gender and cohort year.*

The final four-predictor model explained 39% of variance in TOAFL examination preparedness-a substantive effect size in language education research (Cohen, 1988). Perception of policy relevance emerged as the strongest predictor (β = .42), followed by pesantren background (β = .28), instructional quality (β = .31), availability of practice materials (β = .24), and extrinsic motivational impact (β = .19). These findings confirm that TOAFL preparedness is a multi-determined phenomenon, reflecting the complex interplay of attitudinal, structural, and motivational factors rather than language ability alone.

DISCUSSION

The findings of this study provide partial support for H1 and reinforce the theoretical argument regarding the role of policy perception in shaping student learning behaviours. Students who internalised TOAFL's relevance demonstrated more structured and consistent examination preparation. This finding is consistent with Cheng's (2005) observation that beliefs about an examination's utility mediate between testing policy and the learning strategies students adopt. The pesantren background moderator confirms that a uniform passing threshold risks creating structural inequity-students without equivalent pre-university Arabic language capital are disadvantaged from the outset, a concern compounded by the psychometric evidence that TOAFL items are predominantly calibrated at A2–B1 level while institutional targets aim for B2 competence (Anggara, 2026).

With respect to H2, findings reinforce McNamara and Roever's (2006) argument concerning substantive fairness in standardised testing policy. Disparities in Arabic instructional quality across institutions create conditions in which students do not compete on a level playing field. This pattern parallels that reported by Hamid and Baldauf (2013) in the context of TOEFL requirements at Malaysian universities, where socio-economic capital predicted test outcomes beyond actual linguistic ability. In the TOAFL context, the psychometric findings of Anggara (2026)-demonstrating misalignment between item difficulty and the CEFR B2 target-further complicate this picture: students are not only competing under unequal conditions but are also being assessed by an instrument not yet fully calibrated against the very international standard invoked as its benchmark.

Findings for H3 are theoretically the most nuanced. The dual directional motivational impact replicates the pattern predicted by SDT in high-stakes testing contexts. This mixed washback-increased extrinsic motivation alongside potential erosion of intrinsic motivation-indicates that the TOAFL policy operates as a powerful extrinsic motivator that risks damaging intrinsic motivation when not accompanied by adequate learning support. The finding that pesantren alumni tend to maintain more stable intrinsic motivation suggests that religious and cultural identity mediates motivational orientation towards Arabic (Wahyudi, 2020; Dörnyei, 2009).

The implications of these findings operate at three levels simultaneously. At the instrument level, Anggara's (2026) review provides an empirical mandate for systematic construct reform of TOAFL-from a passive-receptive measurement model towards one that also assesses productive and interactional proficiency. At the institutional level, findings point to the need for bridging policies that connect classroom Arabic curricula to TOAFL competence demands. Without this bridge, students will continue to experience confusion between what is taught and what is tested. At the individual level, successful TOAFL candidates were found to employ eclectic learning strategies-combining formal classroom instruction, independent study, and informal learning communities-with implications for the design of supplementary out-of-class learning programmes.

The sociocultural context of Islamic higher education is distinctive in ways that differentiate it from secular university contexts. Students' integrative motivation-their desire to read Qur'anic texts and classical Islamic scholarship directly-constitutes a motivational capital that, if optimised, can become a powerful driver of Arabic language acquisition. The TOAFL policy should ideally be framed not as an obstacle but as the formalisation of this aspiration. However, for students to perceive it in these terms, institutions need to articulate a more compelling narrative about why Arabic proficiency matters-not merely for passing an examination, but as genuine professional and spiritual preparation.

Several limitations of this study warrant explicit acknowledgement. First, the cross-sectional design precludes strong causal inference; the associations identified require verification through longitudinal investigation. Second, actual TOAFL scores were inaccessible due to institutional privacy policies, so the relationship between questionnaire variables and examination attainment was measured only through self-report. Third, although five institutions were strategically selected, findings may not generalise to private STAI institutions with considerably different resource profiles. Fourth, the Indonesian-language questionnaire may have constrained students' ability to express the full nuance of their experience with Arabic language policy.

Notwithstanding these limitations, the methodological strengths of this study-comprehensive data triangulation between quantitative and qualitative approaches, multi-institutional coverage rare in Indonesian Arabic policy research, and integration with the most recent psychometric review of TOAFL (Anggara, 2026)-yield a richer and more reliable evidential base than the single-site or single-method studies that have previously dominated the TOAFL literature.

A further contribution of this study lies in its application of the Critical Language Testing (CLT) framework to an under-examined non-Western language testing context. While CLT has been extensively applied to English-language testing policy in Southeast Asia (Shohamy, 2001; Hamid & Baldauf, 2013), its application to Arabic language testing in Islamic educational settings remains underdeveloped. The findings of this study suggest that the power dynamics inherent in TOAFL policy-particularly its role in

determining graduation eligibility-create conditions in which students from less privileged Arabic language backgrounds are systematically disadvantaged, irrespective of their actual learning effort or classroom engagement. Addressing this structural inequity requires not only psychometric reform (Anggara, 2026) but also a more deliberate and transparent policy discourse about who the TOAFL is designed to serve and what thresholds can reasonably be set given the diversity of student entry profiles.

The study's integrated findings also have practical implications for the training and professional development of Arabic language instructors at Islamic universities. If instructional quality is the strongest institutional predictor of TOAFL attainment-as demonstrated in the MANOVA results-then investment in instructor capacity building must be treated as a strategic priority rather than a peripheral concern. This includes sustained professional development in communicative Arabic pedagogy, exposure to CEFR-aligned assessment frameworks, and institutional incentives for the development of TOAFL-aligned supplementary teaching materials. Without qualified, well-supported instructors, no policy reform at the examination level alone can close the persistent preparedness gaps documented in this study.

CONCLUSION

This study has generated three principal findings. First, students' perceptions of TOAFL relevance significantly predict their examination preparedness, with pre-university Arabic language educational background (pesantren vs. non-pesantren) serving as an important moderator that reflects linguistic capital disparities that policy design cannot responsibly ignore. Second, barriers related to learning resource accessibility-especially instructional quality and the availability of practice materials aligned with the TOAFL format-significantly affect student outcomes, pointing to a fundamental equity problem in current TOAFL policy implementation. Third, the policy's impact on motivation is characterised by mixed washback: it promotes extrinsic engagement while risking the erosion of intrinsic motivation, particularly among students who experience repeated examination failure.

These findings reinforce and extend the psychometric scholarship demonstrating the need for TOAFL instrument reform to align with international proficiency standards (Anggara, 2026). The student perspectives uncovered in this study provide a policy dimension that has been absent from technical discussions of instrument validity and reliability: a technically valid instrument can still generate harmful policy outcomes if its implementation is not accompanied by a supportive learning ecosystem. In other words, instrument reform alone is insufficient; ecosystem reform must proceed in parallel.

On the basis of these findings, three policy reform agendas are recommended. First, strengthening Arabic language instructional infrastructure equitably across all programmes of study-not only language programmes-including the development of instructional materials that explicitly prepare students for TOAFL format and standards. Second, developing pre-entry diagnostic systems to design differentiated

learning pathways based on students' linguistic profiles, so that not all students must begin from the same starting point towards a uniform target. Third, revising TOAFL passing thresholds in light of the CEFR alignment recommendations proposed by Anggara (2026), accompanied by the development of TOAFL instruments encompassing productive skill dimensions (speaking and writing) to generate a more comprehensive proficiency profile.

At the macro-policy level, this study encourages the Ministry of Religious Affairs and the Ministry of Education, Culture, Research, and Technology to jointly formulate national TOAFL policy guidelines providing minimum institutional benchmarks regarding proportional score thresholds, adequate annual examination frequency, mandatory remedial programme provision, and mechanisms for recognising alternative competency evidence from students with verifiable prior Arabic language education.

Future research agendas should encompass longitudinal studies tracking students from programme entry to graduation to map Arabic language competence development trajectories; comparative cross-institutional analyses incorporating contextual variables more systematically; and the development and validation of TOAFL policy models responsive to the diversity of student profiles and institutional contexts across Indonesian Islamic higher education. Investigation of the role of digital technology and artificial intelligence-enhanced adaptive learning platforms in supporting out-of-class TOAFL preparation is also warranted as part of a more inclusive and sustainable Arabic language learning ecosystem.

In closing, this study affirms that the TOAFL policy-as one manifestation of language-in-education policy-cannot be fairly evaluated through a purely technical psychometric or institutional administrative lens. The voices of students as policy subjects must be an integral component of policy evaluation and improvement cycles. Research that places students' lived experiences, perceptions, and aspirations at the centre of inquiry-as this study has done-is not merely a complement to quantitative analyses; it is an epistemological prerequisite for language education policies that are genuinely fair, effective, and enduring.

ACKNOWLEDGEMENTS

The authors extend their sincere gratitude to all students who participated in this study, to the leadership of the Language Development Centres at the five participating institutions for granting access and institutional approval, and to the anonymous peer reviewers whose critical commentary substantially strengthened the quality of this manuscript. No external funding was received from any governmental, private, or non-profit organisation for this research.

CONFLICT OF INTEREST STATEMENT

The author(s) declare no conflict of interest with respect to the research, authorship, or publication of this article.

REFERENCES

Alderson, J. C. (2004). Assessing reading. Cambridge University Press.

Anggara, S. A. (2026). EVALUASI PSIKOMETRIKA INSTRUMEN TOAFL DALAM BINGKAI CEFR: SEBUAH TINJAUAN LITERATUR SISTEMATIS PADA KEMAHIRAN BERBAHASA ARAB. JIPI (Jurnal Ilmiah Pendidikan Islam), 5(1), 112–122. https://doi.org/10.58788/jipi.v5i1.9535

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. Language Testing, 13(3), 257–279. https://doi.org/10.1177/026553229601300303

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cheng, L. (2005). Changing language teaching through language testing: A washback study. Cambridge University Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment-Companion volume. Council of Europe Publishing.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE Publications.

Deci, E. L., & Ryan, R. M. (2000). The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), Motivation, language identity and the L2 self (pp. 9–42). Multilingual Matters.

Hamid, M. O., & Baldauf, R. B. (2013). Second language errors and features of interlanguage: A study of Bangladeshi university students. Journal of Language Teaching and Research, 4(1), 31–39.

Hidayat, R. (2018). Language standardisation policy at Indonesian Islamic universities: Between the ideal and the real. Jurnal Pendidikan Islam, 7(2), 215–234.

Hughes, A. (2003). Testing for language teachers (2nd ed.). Cambridge University Press.

Kirkpatrick, A. (2011). English as a medium of instruction in Asian education (from primary to tertiary): Implications for local languages and local scholarship. Applied Linguistics Review, 2, 99–120.

Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563–575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Mahyudin, A. (2019). Validity and reliability analysis of the TOAFL instrument at UIN Jakarta. Al-Arabiyah: Jurnal Pendidikan Bahasa Arab, 15(1), 1–18.

McNamara, T., & Roever, C. (2006). Language testing: The social dimension. Blackwell Publishing.

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. Language Learning, 50(1), 57–85. https://doi.org/10.1111/0023-8333.00111

Rosyidi, A. W. (2021). Implementation of the TOAFL policy at UIN Maulana Malik Ibrahim Malang: A case study. Jurnal Bahasa Arab dan Pendidikannya, 4(1), 45–62.

Shohamy, E. (2001). The power of tests: A critical perspective on the uses of language tests. Pearson Education.

Wahyudi, R. (2020). Arabic motivation in Indonesian Islamic higher education: A mixed-methods exploration. International Journal of Arabic Language Teaching, 2(1), 1–22.

Wall, D. (2005). The impact of high-stakes testing on classroom teaching: A case study using insights from testing and innovation theory. Cambridge University Press.

Zulhannan. (2014). Teknik pembelajaran bahasa Arab interaktif [Interactive Arabic language teaching techniques]. PT RajaGrafindo Persada.