

An Argument-Based Validation Framework for Automated Scoring of Higher-Order Thinking in Arabic Essays

Rozaanah Rozaanah¹⁾, Slamet Riyadi²⁾

^{1,2)}Sekolah Tinggi Agama Islam As-sunnah, Indonesia

E-mail: ¹⁾rozaanah@assunnah.ac.id, ²⁾slametriyadi@assunnah.ac.id

 <https://doi.org/10.54214/alfawaid.Vol16.Iss2.1052>

ABSTRACT

Automated essay scoring (AES) for Arabic has advanced quickly, but most studies report agreement statistics without stating what an automated score is meant to show or which uses it can support. The gap is sharper for higher-order thinking skills (HOTS), where good answers legitimately diverge. This article develops an argument-based validation framework for Tahkīm, a semantic-similarity engine designed to score Arabic essays written by Indonesian undergraduates at the analysing, evaluating, and creating levels of the revised Bloom taxonomy. Using a conceptual design approach, the study specifies the intended interpretation and uses of Tahkīm scores, derives an interpretation-use argument from Kane's framework and the Williamson-Xi-Breyer criteria, and synthesizes Arabic AES research to identify threats that arise in Arabic. The framework comprises four inferences (scoring, generalization, extrapolation, and decision), each with a warrant, testable assumptions, required evidence, and a decision rule fixed before data collection. It identifies eight threats, among them orthographic normalization, colloquial vocabulary, incomplete anchor coverage for divergent creating-level responses, and the confounding of reasoning with language proficiency, and pairs each with a design response and an empirical check. A staged evaluation plan sets the minimum evidence for each permitted use, from formative feedback to first-pass scoring under lecturer review, and excludes unsupervised grading until all four inferences hold. The framework gives Arabic departments in Islamic higher education a procedure for stating and testing claims about automated scores before adopting them, and it is the protocol for the empirical validation of Tahkīm.

Keywords : Arabic as a Foreign Language, Argument-Based Validity, Automated Essay Scoring, Higher-Order Thinking Skills, Validation Design

INTRODUCTION

Lecturers who teach Arabic writing (*insyāʾ*) in Indonesian Islamic colleges spend much of each semester marking essays, and the essays that matter most for learning take the longest to mark. These are responses in which students analyse a claim, weigh evidence, or build an argument of their own. They correspond to the analysing, evaluating, and creating levels of the revised Bloom taxonomy (Anderson & Krathwohl, 2001; Krathwohl, 2002), and they produce answers that differ legitimately from one another. Two competent students can reach opposite conclusions and both

deserve a high score. Human raters handle this divergence imperfectly, and fatigue and drift over a long marking session make the problem worse (Weigle, 2002).

Automated essay scoring (AES) has been offered as a partial answer to this burden for decades. Early systems such as Project Essay Grade and e-rater predicted scores from surface features such as length, vocabulary, and error counts (Page, 1966; Attali & Burstein, 2006). Neural models later learned representations directly from text (Taghipour & Ng, 2016; Dong et al., 2017), and transformer encoders such as BERT raised agreement with human raters further (Devlin et al., 2019; Beseiso & Alzahrani, 2020; Beseiso et al., 2021; Wang et al., 2022). More recent studies test large language models as raters of second-language writing and report reliability that varies with task and prompt design (Mizumoto & Eguchi, 2023; Yancey et al., 2023; Pack et al., 2024). Two systematic reviews trace this development, and both show how heavily the field depends on English datasets and English benchmarks (Ke & Ng, 2019; Ramesh & Sanampudi, 2022).

The methods fall into three broad families (Ramesh & Sanampudi, 2022). Feature-based models predict scores from engineered indicators. Fine-tuned transformer models learn the mapping from text to score from large numbers of rated essays. Similarity-based methods compare a new response with reference answers whose scores are known and need only a few scored anchors per prompt (Shehab et al., 2018). The third family appeals to small programs, because a lecturer can add a new prompt by writing references instead of collecting and rating hundreds of essays. Its weakness follows from the same design. A similarity engine rewards resemblance to its anchors, so its accuracy depends on how well the anchors cover the range of acceptable answers, and that dependence should matter most when the construct is open-ended.

Arabic raises problems that English-trained methods do not anticipate. Its morphology packs several morphemes into one orthographic word; its spelling varies in the seat of the hamzah and in the writing of *tā' marbūṭah* and *alif maqṣūrah*; and learners move between Modern Standard Arabic and colloquial forms. Arabic pre-trained language models narrowed part of this gap. AraBERT outperformed multilingual BERT on most of the Arabic tasks its developers tested (Antoun et al., 2020), and ARBERT and MARBERT extended coverage to dialectal text (Abdul-Mageed et al., 2021). Arabic essay scoring followed. Shehab et al. (2018) combined string-based and corpus-based similarity measures to grade Arabic essays. Gaheen et al. (2021) trained a neural network optimized with an elitist Jaya algorithm to score students' Arabic essays within a personalized system of instruction. Ghazawi and Simpson (2024) released the AR-AES dataset with a BERT-based baseline, and in a follow-up study they reported that a fine-tuned AraBERT model reached a QWK

of 0.88, while human inter-rater agreement across questions ranged from 0.78 to 0.90 (Ghazawi & Simpson, 2025). In Indonesia, Susilawati et al. (2025) developed an automated short-essay assessment model for an Islamic education study program and judged it valid on the basis of expert review and user trials.

These studies show that Arabic essays can be scored automatically with useful accuracy. Most of them, however, report accuracy alone: agreement or error statistics computed over a dataset. A scoping review of Arabic AES documents the rapid growth of models and datasets (Elhaddadi et al., 2025), and a systematic review of 31 deep-learning AES studies relevant to Islamic education, published between 2020 and 2025, found that the literature concentrates on how closely algorithms reproduce human scores (Putri et al., 2025). Relatively few of the studies covered by these reviews state what an automated score is meant to show and which decisions it may inform, and relatively few isolate higher-order thinking as the construct being scored. For HOTS, the omission has a concrete cost. A high QWK computed over a mixed set of tasks can conceal weak agreement on the open-ended prompts that HOTS assessment is designed to include, and an agreement figure alone cannot tell a department whether a tool is fit for feedback, for grading, or for neither.

Validity theory offers a way to organize the missing evidence. Kane (1992, 2013) treats validation as the evaluation of an interpretation-use argument, a chain of inferences that runs from an observed performance to a score (scoring), from that score to performance on comparable tasks and raters (generalization), from there to the target ability (extrapolation), and finally to the decisions the score will inform (decision). Each inference rests on a warrant and needs its own evidence, and a weak link limits the whole argument (Cook et al., 2015). Williamson et al. (2012) turned this reasoning into operational criteria for automated scoring. Their guidelines ask for machine-human agreement of at least 0.70 in QWK or correlation, and they add a degradation criterion: machine-human agreement should not fall more than 0.10 below human-human agreement. Other work in this tradition adds that validation must also address fairness across subgroups and the consequences of using the scores (Xi, 2010; Ramineni & Williamson, 2013).

Agreement statistics say little about how students and lecturers respond to automated scores. Nazaretsky et al. (2026) asked 472 students to evaluate teacher and AI feedback on their own writing; the students rated AI feedback lower once they knew its source and judged it less genuine. Acceptance models such as UTAUT link the intention to use a technology with expected performance, expected effort, and social influence (Venkatesh et al., 2003, 2012), and a systematic review confirms that these constructs predict adoption in higher education (Xue et al., 2024).

Kasneci et al. (2023) add that language models in education raise questions of bias and accountability that accuracy figures cannot settle. In Kane's terms, these questions belong to the decision inference. Formative use adds one further condition: learners can use feedback to regulate their own writing only when they understand it (Zimmerman, 2002; Panadero, 2017).

Indonesia is an informative setting for these questions. It has the largest Muslim population of any country (Pew Research Center, 2025). Civil registration data released in March 2026 put its population at 288,315,089 at the end of 2025, of whom 87.15% (about 251.3 million) were Muslim (Direktorat Jenderal Kependudukan dan Pencatatan Sipil, 2026). Arabic is a compulsory subject throughout its madrasah system (Kementerian Agama Republik Indonesia, 2019) and a core subject in its Islamic higher education institutions (PTKI), whose Arabic departments train the teachers who will later assess Arabic writing in madrasahs. Taḥkīm is being developed at STAI As-Sunnah Deli Serdang, a private Islamic college in North Sumatra with a residential *ma'had*, to score Arabic essays against a HOTS rubric. Before such a tool is used with students, the claims its scores are meant to support have to be stated and the evidence needed for each claim specified. Kane (2013) recommends this order: the interpretation-use argument comes first, and evidence is then gathered to test its most doubtful assumptions. Relatively few Arabic AES studies make this step explicit, and none of those reviewed above does so for HOTS. This article takes that step for Taḥkīm.

A typical prompt from the *insyā* 'course illustrates the kind of task Taḥkīm is meant to score:

حَلِّلْ وَقَيِّمْ: هَلْ يُمَكِّنُ لِلتَّقْوِيمِ الْآلِيَّ أَنْ يَحُلَّ مَحَلَّ الْمُعَلِّمِ فِي تَصْحِيحِ الْمَقَالَاتِ؟ اذْكُرْ حُجَجَكَ.

("Analyze and evaluate: can automated assessment replace the teacher in marking essays? State your arguments.") Prompts of this kind target the upper levels of the taxonomy and invite divergent answers, which is why they test an automated scorer most severely. The article addresses three design questions:

This study examines the interpretations, uses, inferences, and assumptions required to establish a validity argument for Taḥkīm's higher-order thinking skills (HOTS) scores. It also identifies potential threats to the validity argument arising from the use of semantic similarity to score Arabic essays and explores how Taḥkīm's design can address these threats. Furthermore, the study establishes the evidence and predetermined decision rules necessary to justify each intended use of Taḥkīm's scores.

METHOD

Research Design

This is a conceptual design study. Its products are a validation framework and an evaluation protocol, and it reports no empirical estimates of Tahkīm's performance. The approach follows Kane's (2013) advice to specify the interpretation-use argument before evidence is gathered, and the practical guide of Cook et al. (2015), which treats validation as the prioritized testing of the assumptions in that argument. The work proceeded in five steps. The first step specified the intended interpretation of Tahkīm scores and the uses they are meant to serve in the program. The second derived the inferences and warrants of the argument from Kane (1992, 2013) and translated them into the operational criteria of Williamson et al. (2012). The third synthesized Arabic AES research to identify threats that arise in Arabic and in HOTS scoring. The fourth linked each design decision in Tahkīm to an assumption in the argument, so that every technical choice has a corresponding check. The fifth fixed the evidence requirements and decision rules for each inference and arranged the evidence collection into stages.

Sources for the Synthesis

The synthesis was purposive rather than systematic. It started from two recent reviews, a scoping review of Arabic AES (Elhaddadi et al., 2025) and a systematic review of deep-learning AES in Islamic education (Putri et al., 2025), together with the broader review by Ramesh and Sanampudi (2022). Primary studies of Arabic essay scoring were then examined for the threats they reported or implied (Shehab et al., 2018; Gaheen et al., 2021; Ghazawi & Simpson, 2024, 2025; Susilawati et al., 2025), as were studies of Arabic language models (Antoun et al., 2020; Abdul-Mageed et al., 2021). The measurement literature on automated scoring supplied the evaluation criteria (Xi, 2010; Williamson et al., 2012; Ramineni & Williamson, 2013). Because the synthesis was not exhaustive, the list of threats presented below is a starting set that later studies can extend.

Context of Intended Use

Tahkīm is intended for the Arabic Language Education (Pendidikan Bahasa Arab, PBA) program at STAI As-Sunnah Deli Serdang, North Sumatra, Indonesia. Students in the program learn Arabic as a foreign language and live in a residential *ma'had* where Arabic is part of daily routines. The target tasks are essay assignments in *insyā'* (composition) courses that ask students to analyse, evaluate, or create in Arabic. Lecturers [currently mark these essays by hand with a HOTS rubric of four levels]. The framework assumes that the same rubric will serve as the criterion for evaluating Tahkīm, so that machine and human scores share one scale.

The Taḥkīm Scoring Engine

Taḥkīm scores an essay by comparing its meaning with reference responses whose scores are already known. It does not require a large set of previously scored essays for training, which suits Arabic departments where such datasets do not exist. The engine is designed to work in four steps. The first step is text normalization. Taḥkīm unifies orthographic variants that carry no difference in meaning for scoring purposes: it maps the hamzated forms of alif (ā, ʾ, ʾ) to bare alif (a), treats final ع and ى as one letter, maps ّ to َ, removes short-vowel diacritics and *tatwīl*, and deletes any sentence copied verbatim from the prompt [confirm the rules implemented]. The second step is encoding. Each essay and each reference response is converted into a single vector by [model and checkpoint, e.g., AraBERT v2 (aubmindlab/bert-base-arabertv02) with mean pooling]. Encoders of this kind, trained in the Sentence-BERT manner, place texts with similar meaning close together, so that two texts can be compared by the cosine of the angle between their vectors (Reimers & Gurevych, 2019). Because such encoders accept at most 512 tokens, longer essays are split into overlapping segments whose vectors are averaged.

The third step is the construction of reference responses. For each prompt, lecturers write at least three reference responses at each level of the four-level HOTS rubric, and the reference responses at the same level are written to follow different lines of argument. A panel of raters confirms the level of every reference before it is used. Level-4 references model the analysis, evaluation, or original proposal that the prompt asks for; level-1 references restate the prompt or list facts without reasoning.

The fourth step maps similarity to a score. Taḥkīm computes the cosine similarity between the student essay and every anchor for that prompt, averages the similarities within each rubric level, and assigns the level with the highest mean similarity. The output is an integer from 1 to 4 on the scale the human raters use. Taḥkīm also reports the highest single similarity value; an essay whose closest anchor falls below a threshold set during piloting is flagged for lecturer review instead of receiving an automatic score.

RESULT AND DISCUSSION

Intended Interpretation and Uses of Taḥkīm Scores

A Taḥkīm score is interpreted as the level of higher-order thinking that one Arabic essay demonstrates in response to one prompt, judged against the program's HOTS rubric. The claim is bounded in three ways. It concerns the essay and says nothing directly about a student's general

ability. It concerns reasoning as the rubric describes it, and it excludes grammatical and orthographic accuracy, which Taḥkīm's normalization step removes from view. It also holds only for prompts that have validated anchors, so a new prompt has no Taḥkīm score until its anchors are written and checked.

Table 1 lists the uses the framework considers, in increasing order of stakes. Each use needs more of the argument to hold than the one before it. Formative feedback on drafts requires the scoring inference, because a student acts on the score only as a hint about where the argument is weak. First-pass scoring, in which a lecturer confirms or corrects every score, requires the scoring and generalization inferences for the cognitive level concerned. Use of Taḥkīm scores in course grades without lecturer review requires all four inferences, together with evidence that no group of students is disadvantaged. The framework excludes placement and ranking decisions altogether, since the claim concerns single essays.

Table 1. Intended uses of Taḥkīm scores and the inferences each use requires

Use	Stakes	Inferences that must be supported	Status in the framework
Formative feedback on drafts	Low	Scoring	Permitted once the scoring rule is met for the level
First-pass scoring confirmed or corrected by a lecturer	Moderate	Scoring and generalization for the level concerned	Permitted once both rules are met for the level
Course grading without lecturer review	High	Scoring, generalization, extrapolation, and decision, including subgroup checks	Excluded until all rules are met in two consecutive semesters
Placement or ranking of students	High	A different claim about student ability	Outside the framework

The Interpretation-Use Argument

Table 2 sets out the argument. For each inference it states the warrant, the assumptions on which the warrant depends, the evidence that would back those assumptions, and the most plausible rebuttal. Stating the rebuttal in advance matters, because it tells the evaluator which result would count against the claim (Cook et al., 2015).

Table 2. Interpretation-use argument for Tahkīm scores

Inference	Warrant	Key assumptions	Backing required	Plausible rebuttal
Scoring	Tahkīm assigns the rubric level that trained raters assign to the same essay	Anchors represent each level; embeddings reflect reasoning and not only topic words; the human criterion is reliable	Machine-human agreement overall and by level; review of disagreements	Agreement rests on vocabulary shared with the anchors
Generalization	A score on one prompt would be similar on a comparable prompt at the same level	Anchor sets are of similar quality across prompts; human raters agree adequately	Agreement by prompt, with at least two prompts per level; human-human agreement before adjudication	Agreement is good on one prompt and weak on another at the same level
Extrapolation	Scores reflect higher-order thinking in Arabic writing beyond the specific task	Scores relate to independent writing measures but are not reducible to proficiency, length, or copying	Correlations with external measures, with essay length, and with prompt-copy rate	Scores track general proficiency or essay length
Decision	Using the scores for the permitted uses has acceptable consequences	Users understand the scores; no subgroup is disadvantaged; lecturer review is real	Questionnaire and interviews; subgroup score differences; audit of lecturer overrides	Students who use colloquial forms are scored lower; lecturers accept machine scores without checking

The scoring inference carries the heaviest load, because every later inference assumes it. Its first assumption, that the anchors represent each level, depends on the people who write them. If all level-4 anchors follow one line of argument, a strong essay that argues the opposite position will look distant from level 4. The design response, several anchors per level written along different lines, is cheap, but its effect has to be measured. The second assumption is harder to secure. Sentence encoders place texts on the same topic close together whether or not they reason well, so high agreement can arise because weak and strong essays differ in vocabulary rather than in argument.

Reviewing the essays on which Taḥkīm and the raters disagree, and checking whether they share topic words with the anchors, is the most direct test of this rebuttal. The third assumption concerns the criterion itself. Human raters of HOTS essays disagree for legitimate reasons (Weigle, 2002), and a machine cannot be expected to agree with a criterion that does not agree with itself.

Generalization asks whether a result on one prompt would hold on another. Ghazawi and Simpson (2025) found that agreement between automated and human scores varied considerably from one question to another within a single Arabic dataset, so an overall figure cannot stand in for each prompt. The framework therefore requires at least two prompts per cognitive level; with one prompt per level, the effect of the level cannot be separated from the effect of the particular prompt. Extrapolation asks whether the scores mean more than performance on the task at hand. Correlations with independent measures of Arabic writing should be positive, but they should also stay clearly below the agreement with the human HOTS score. If Taḥkīm scores tracked general proficiency almost perfectly, they would be measuring proficiency and not the reasoning that the rubric describes.

The decision inference concerns consequences, and the studies of trust reviewed in the introduction become part of the argument here. If students judge AI feedback less genuine once they know its source (Nazaretsky et al., 2026), they may discount Taḥkīm's feedback even when it is accurate, and the questionnaire has to detect that. If accountability for AI output has to rest with people (Kasneci et al., 2023), first-pass scoring meets that requirement only if lecturers actually check the scores. An audit of how often lecturers change Taḥkīm's scores provides the evidence: an override rate near zero on essays where the raters themselves disagree would suggest that review has become a formality. For formative use, a further condition applies. Feedback supports self-regulated revision only when learners can connect it to what they should do next (Zimmerman, 2002; Panadero, 2017), so a bare level from 1 to 4 is a weak form of feedback. Score reports should name the rubric descriptor that the essay met and the one it did not.

Threats Specific to Arabic and Design Responses

The synthesis identified eight threats that a semantic-similarity engine faces when it scores HOTS in Arabic essays. Table 3 lists each threat, how it could distort scores, the design response built into Taḥkīm, and the check that the evaluation will apply.

Table 3. Threats to the validity of Tahkīm scores, design responses, and checks

Threat	Possible distortion	Design response	Check
T1. Orthographic variation (hamzah, <i>tā'</i> <i>marbūṭah</i> , <i>alif maqṣūrah</i>)	Spelling differences shift embeddings; raters may penalize spelling despite the rubric	Normalization before encoding; rubric excludes spelling	Score a sample in raw and normalized form; code disagreements for spelling density
T2. Colloquial vocabulary and constructions	Essays using colloquial forms lie farther from anchors written in Modern Standard Arabic	Encoder with dialect coverage considered (Abdul-Mageed et al., 2021); essays with frequent colloquial forms flagged for review	Score differences between students with high and low colloquial use
T3. Incomplete anchor coverage for divergent responses	Original, well-argued essays, especially at C6, receive low scores	At least three anchors per level along different lines of argument; low maximum similarity triggers lecturer review	Agreement by level; review of flagged essays
T4. Confounding of reasoning with proficiency	Fluent but shallow essays resemble high-level anchors in language	Anchors at the same level vary in language quality	Correlation with the proficiency test compared with correlation with the human HOTS score
T5. Essay length	Longer essays accumulate more overlap with anchors	Mean-pooled vectors, which do not grow with length	Correlation of length with Tahkīm scores compared with its correlation with human scores
T6. Copying from the prompt	Reproduced prompt text raises similarity without reasoning	Sentences copied from the prompt deleted before encoding	Correlation of copy rate with Tahkīm scores compared with human scores
T7. Input length limit of the encoder	Truncated essays lose their conclusions, where evaluation often appears	Overlapping segments averaged	Compare segmented and truncated scores on long essays
T8. Weak human criterion	Machine agreement is judged against an unreliable standard	Rater calibration, blind double rating, adjudication rule fixed in advance	Human-human agreement before adjudication must meet its own threshold

Three of these threats need further discussion. The first is normalization (T1), because it encodes a decision about the construct. By mapping ڤ to ڤ and unifying the forms of alif, Taḥkīm ignores spelling errors that a lecturer marking language accuracy would penalize. For a HOTS rubric the decision is defensible, since spelling variation is construct-irrelevant when the construct is reasoning (Messick, 1995). It also means that Taḥkīm scores carry no information about orthographic accuracy, and a department that reports them should say so and assess accuracy separately. Human raters may be swayed by spelling even when the rubric tells them to disregard it, so part of any machine-human disagreement could originate with the raters. Coding a sample of disagreements for spelling density would show which side the error lies on.

The second is the confounding of reasoning with proficiency (T4). Indonesian learners of Arabic differ widely in their prior schooling; students who did not attend a pesantren often begin with a gap in Arabic competence relative to their pesantren-educated classmates (Batalipu et al., 2026). A student with years of pesantren education may write fluent Arabic that says little, while a student with less exposure may build a sound argument in simple sentences. Cumming (1989) observed this separation among adult writers of English as a second language: writing expertise and second-language proficiency made distinct contributions to the quality of their compositions. A similarity engine compares whole texts, so fluency can pull an essay toward the high anchors.

Writing anchors at the same level in varied language quality weakens this pull, and the extrapolation check tests whether it has been weakened enough. The third is anchor coverage at the creating level (T3). A creating prompt asks for a proposal the anchor writers may not have foreseen. The flag for low maximum similarity turns this weakness into a routing rule: essays unlike anything Taḥkīm has seen go to a lecturer. Ghazawi and Simpson (2025) reported that large language models reached lower agreement than fine-tuned AraBERT on their Arabic dataset, so replacing the similarity engine with a general-purpose language model would not remove the need for these checks.

Evidence Plan and Decision Rules

Table 4 translates the argument into an evaluation plan. Each row names the evidence, how it will be collected, the statistic that summarizes it, and the rule that decides whether the corresponding assumption holds. The rules are fixed before any data are collected, so that the evaluation cannot adjust its standards to its results.

Table 4. Evidence, statistics, and decision rules for the evaluation of Tahkīm

Evidence	Collection	Statistic	Decision rule
Human criterion	Two raters score every essay blind after calibration; essays differing by more than one point are adjudicated	QWK between raters before adjudication, with 95% CI	At least 0.70; otherwise recalibrate before evaluating Tahkīm
Overall machine-human agreement	Tahkīm scores for all essays	QWK, Pearson and Spearman correlations, exact and adjacent agreement; 95% CI by student-level bootstrap	QWK and r of at least 0.70, with the lower CI bound reported
Degradation	As above, against the resolved score and against each rater	Human-human QWK minus machine-human QWK	No more than 0.10
Agreement by level and prompt	At least two prompts per cognitive level	QWK per level and per prompt, with 95% CI	A level is released for use only if its QWK is at least 0.70
Score scale	All essays	Standardized mean difference, Tahkīm minus human	Absolute value no more than 0.15
Subgroups	Proficiency bands; high versus low colloquial use	Standardized mean difference within each group	Absolute value no more than 0.10
Extrapolation	Final <i>insyā</i> 'grade from an earlier semester; institutional Arabic test	Pearson r with 95% CI	Positive, CI excludes zero, and clearly below the machine-human r
Length copying and	Word count; proportion of text copied from the prompt	Correlation with Tahkīm scores and with human scores	Tahkīm's correlation no higher than the human correlation
Consequences	Questionnaire on objectivity, usefulness, genuineness, source credibility, and performance expectancy; interviews; override audit	Means, SDs, and Cronbach's alpha; reflexive thematic analysis; override rate	Judged against each use; no fixed numerical threshold

The agreement thresholds come from Williamson et al. (2012): a QWK or correlation of at least 0.70, a degradation of no more than 0.10, and standardized mean differences of no more than 0.15 overall and 0.10 for subgroups. Kappa values will be described with the Landis and Koch (1977) labels, which are conventions rather than standards. QWK follows Cohen (1968). Confidence

Published by STAI Ali bin Abi Thalib Surabaya

Website Jurnal : <https://jurnal.stai-ali.ac.id/index.php/Alfawaid>

DOI Jurnal : <https://doi.org/10.54214/alfawaid>

intervals will be obtained by bootstrap resampling of students, so that all essays by one student stay together; resampling essays as if they were independent would understate the uncertainty. Human-human agreement will be computed from the two raters' independent scores before adjudication, and Taḥkīm will be compared both with the resolved score and with each rater, because the resolved score is more reliable than a single rating and would otherwise flatter the machine.

The plan also sets the scale of the study. A pipeline test on simulated data, generated only to check the analysis code, used 60 students who each wrote one essay at each of the three levels. In that test the 95% confidence interval for overall QWK was about 0.13 wide, and the interval for a single level ranged from about 0.13 to 0.25 wide. Intervals of that width are too broad for a decision about one level based on one prompt. The empirical study should therefore include at least 60 students and at least two prompts per level. The qualitative strand will follow an explanatory sequential logic (Creswell & Plano Clark, 2018): students from each proficiency band and the participating lecturers will be interviewed after the quantitative results are known, the transcripts will be analysed with reflexive thematic analysis (Braun & Clarke, 2006), and a joint display will set each inference against its quantitative and qualitative evidence.

Staged Evaluation

The evidence will be collected in four stages, and each stage releases a use only when its rules are met. In the first stage, lecturers write and cross-check the anchor sets, and a panel of experts rates the relevance of the rubric descriptors and the level of each anchor. In the second stage, a small pilot tests the rating procedure, the analysis pipeline, and the similarity threshold for flagging essays, with no consequences for students. The third stage is the main validation study described in Table 4, carried out under the ethical procedures of the institution and with students' informed consent. If the scoring rule is met for a level, Taḥkīm may be used for formative feedback at that level; if the generalization rule is also met, it may be used for first-pass scoring under lecturer review. The fourth stage is monitoring. Agreement, subgroup differences, and override rates are recomputed each semester, since prompts, anchors, and student intakes change, and a validity argument for automated scores has to be renewed when they do (Williamson et al., 2012).

Contribution in Relation to Previous Work

The framework differs from earlier Arabic AES work in what it treats as the result. Ghazawi and Simpson (2024, 2025) evaluated models on a shared dataset and reported agreement; Susilawati et al. (2025) established the validity of an automated short-essay model through expert judgement and user trials. Both kinds of evidence belong in a validity argument, but neither says which uses of

a score the evidence supports. The framework presented here starts from the uses, states the assumptions each use depends on, and fixes the evidence and the thresholds before any data exist. Fixing the thresholds first keeps the evaluators from deciding what counts as good enough after they have seen the numbers.

The framework also makes the Arabic-specific threats explicit. Orthographic normalization, colloquial usage, and the gap between fluency and reasoning are familiar to any lecturer who marks Arabic essays by Indonesian students, yet they rarely appear as named threats in AES evaluations. Once they are named as threats, a department can test them as assumptions instead of leaving them to the judgement of individual lecturers. Most of the framework does not depend on the similarity engine. A department considering a large language model as a rater, as tested for second-language English writing by Mizumoto and Eguchi (2023), Yancey et al. (2023), and Pack et al. (2024), would face the same inferences and most of the same threats, although T3, T5, and T6 would take different forms. Most published AES evidence comes from English, where large scored datasets exist (Ramesh & Sanampudi, 2022). A procedure that needs only expert anchors, trained raters, and an Arabic encoder is within reach of Arabic departments in PTKI and of programs teaching other languages with few scored essay datasets.

CONCLUSIONS

This article set out an argument-based validation framework for Taḥkīm, a semantic-similarity engine designed to score higher-order thinking in Arabic essays. The framework limits the claim to the level of reasoning shown in one essay on one prompt, distinguishes three uses of increasing stakes, and specifies four inferences with their assumptions, backing, and rebuttals. It names eight threats that arise when Arabic essays are scored by similarity, pairs each with a design response and a check, and fixes decision rules based on the criteria of Williamson et al. (2012) before any data are collected. Under the framework, Taḥkīm may be used for formative feedback and then for first-pass scoring under lecturer review, level by level, as the evidence permits. Unsupervised grading remains excluded until all four inferences hold.

For Arabic language education, the contribution is practical: departments in Islamic higher education that consider automated scoring now have a procedure for stating what the scores are meant to show, for deciding in advance what evidence would justify each use, and for testing the assumptions most likely to fail in Arabic. The framework also gives the empirical validation of Taḥkīm a protocol that reviewers and readers can check against its eventual results.

The framework has limitations. It rests on a purposive synthesis, and a systematic search could add threats it does not list. Its thresholds come from large-scale testing programs in English and may need adjustment for formative classroom use in Arabic. It was developed for one program with a residential *ma'had*, whose students have more daily contact with Arabic than many PTKI students. It has not yet been reviewed by external experts in measurement or Arabic NLP. Its largest limitation is that it contains no evidence about Taḥkīm's actual performance, so its value depends on the validation study it prepares.

Three steps should follow. The framework should be reviewed by an external panel before data collection. The validation study should then be carried out as specified, with at least 60 students and two prompts per level, and its results reported against the rules in Table 4, including rules that are not met. Finally, the framework should be applied in other PTKI Arabic programs, with and without residential components, to test whether its threats and thresholds hold beyond the setting for which it was written.

REFERENCES

- Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2021). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 7088–7105). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.551>
- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools* (pp. 9–15). European Language Resources Association.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning and Assessment*, 4(3), 1–30.
- Batalipu, A., Miolo, M. I., & Sarif, S. (2026). Learning Arabic in heterogeneous classrooms: Experiences of non-pesantren students in Indonesian Islamic higher education. *Al-Ta'rib: Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab UIN Palangka Raya*, 14(1), 81–96. <https://doi.org/10.23971/altarib.v14i1.11341>

- Beseiso, M., & Alzahrani, S. (2020). An empirical analysis of BERT embedding for automated essay scoring. *International Journal of Advanced Computer Science and Applications*, 11(10), 204–210. <https://doi.org/10.14569/IJACSA.2020.0111027>
- Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, 33(3), 727–746. <https://doi.org/10.1007/s12528-021-09283-1>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4), 213–220. <https://doi.org/10.1037/h0026256>
- Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, 49(6), 560–575. <https://doi.org/10.1111/medu.12678>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
- Cumming, A. (1989). Writing expertise and second-language proficiency. *Language Learning*, 39(1), 81–135. <https://doi.org/10.1111/j.1467-1770.1989.tb00592.x>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>
- Direktorat Jenderal Kependudukan dan Pencatatan Sipil. (2026). *Data kependudukan bersih semester II tahun 2025* [Data set]. Kementerian Dalam Negeri Republik Indonesia.
- Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning* (pp. 153–162). Association for Computational Linguistics. <https://doi.org/10.18653/v1/K17-1017>
- Elhaddadi, C., Zeroual, I., & El Kah, A. (2025). Automatic Arabic essays scoring: A scoping review. In B. Hdioud & S. L. Aouragh (Eds.), *Arabic language processing: From theory to practice* (Communications in Computer and Information Science, Vol. 2339, pp. 45–60). Springer. https://doi.org/10.1007/978-3-031-79164-2_4

- Gaheen, M. M., ElEraky, R. M., & Ewees, A. A. (2021). Automated students Arabic essay scoring using trained neural network by e-Jaya optimization to support personalized system of instruction. *Education and Information Technologies*, 26(1), 1165–1181. <https://doi.org/10.1007/s10639-020-10300-6>
- Ghazawi, R., & Simpson, E. (2024). *Automated essay scoring in Arabic: A dataset and analysis of a BERT-based system* (arXiv:2407.11212). arXiv. <https://doi.org/10.48550/arXiv.2407.11212>
- Ghazawi, R., & Simpson, E. (2025). How well can LLMs grade essays in Arabic? *Computers and Education: Artificial Intelligence*, 9, Article 100449. <https://doi.org/10.1016/j.caeai.2025.100449>
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527–535. <https://doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. 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>
- Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6300–6308). IJCAI. <https://doi.org/10.24963/ijcai.2019/879>
- Kementerian Agama Republik Indonesia. (2019). *Keputusan Menteri Agama Nomor 183 Tahun 2019 tentang Kurikulum Pendidikan Agama Islam dan Bahasa Arab pada Madrasah*. Kementerian Agama Republik Indonesia.
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>

- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), Article 100050. <https://doi.org/10.1016/j.rmal.2023.100050>
- Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2026). Who gives feedback matters: Student biases towards human and AI-generated formative feedback. *Journal of Computer Assisted Learning*, 42(1), Article e70153. <https://doi.org/10.1111/jcal.70153>
- Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6, Article 100234. <https://doi.org/10.1016/j.caeai.2024.100234>
- Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5), 238–243.
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, Article 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Pew Research Center. (2025, June 9). *How the global religious landscape changed from 2010 to 2020*. <https://www.pewresearch.org/religion/2025/06/09/how-the-global-religious-landscape-changed-from-2010-to-2020/>
- Putri, R. K. A., Kusaeri, K., & Suparto, S. (2025). Deep learning in automated essay scoring for Islamic education: A systematic review. *Online Learning in Educational Research*, 5(2), 319–338. <https://doi.org/10.58524/oler.v5i2.753>
- Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, 55(3), 2495–2527. <https://doi.org/10.1007/s10462-021-10068-2>
- Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1), 25–39. <https://doi.org/10.1016/j.asw.2012.10.004>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1410>
- Shehab, A., Faroun, M., & Rashad, M. (2018). An automatic Arabic essay grading system based on text similarity algorithms. *International Journal of Advanced Computer Science and Applications*, 9(3), 263–268. <https://doi.org/10.14569/IJACSA.2018.090337>

- Susilawati, E., Siregar, E. S., Ekalestari, S., & Harahap, E. S. (2025). Development of a digital assessment model for automated short essay scoring in Islamic education study program. *Fitrah: Journal of Islamic Education*, 6(1), 208–223. <https://doi.org/10.53802/fitrah.v6i1.1195>
- Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891). Association for Computational Linguistics. <https://doi.org/10.18653/v1/D16-1193>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. <https://doi.org/10.2307/30036540>
- Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36(1), 157–178. <https://doi.org/10.2307/41410412>
- Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3416–3425). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.249>
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511732997>
- Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1), 2–13. <https://doi.org/10.1111/j.1745-3992.2011.00223.x>
- Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3), 291–300. <https://doi.org/10.1177/0265532210364643>
- Xue, L., Rashid, A. M., & Ouyang, S. (2024). The unified theory of acceptance and use of technology (UTAUT) in higher education: A systematic review. *SAGE Open*, 14(1). <https://doi.org/10.1177/21582440241229570>
- Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 576–584). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.bea-1.49>

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2