6th CELTI
Conference in English Language Teaching
AN INTERNATIONAL VIRTUAL CONFERENCE
Decolonize English at the Age of
Artificial Intelligence: Rethinking
Language, Literature, and Pedagogy

The Hidden Bias Of Artificial Intelligence In Indonesian Elt Assessment: A Mixed-Methods Study Of Grammarly And Turnitin Scores On Student Writing

Author:
I Putu Andre Suhardiana*

Affiliation:
Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar, Indonesia

Corresponding email:
*putuandresuhardiana@gmail.com**

DOI:
https://doi.org/10.24090/zdy23088

Pages:
188-200

Abstract: The growing use of AI tools in ELT assessment, like Grammarly and Turnitin, poses risks of algorithmic bias against non-native varieties of English. Although there is extensive research to show bias against Indian, Nigerian and Chinese English writing, no empirical studies about Indonesian English have been investigated nor isolated, comparing AI scores versus human rubric marking which allows culturally situated features like code-switching, politeness markers and more indirect rhetorical structures. This sequential explanatory mixed-methods study examined how Indonesian EFL writing features that are grounded in local culture were penalized by Grammarly and Turnitin, and what lecturers perceived about this bias. Presumably, Grammarly, Turnitin and the three lecturers were invited to score sixteen essays by six university students, then subsequently interviewed semi-structured style, a qualitative design. In the quantitative data, some strong negative bias was revealed: essays were penalized by Grammarly for larger -11.8 points (93.75%) of essays or Turnitin scores -6.9 points (87.5% of essays). The harshest penalties were levied for code-switching (Grammarly bias -15.2), while the moderate ones were reserved for politeness markers (-12.6) and indirect rhetorical structures (-8.9). Lecturers were in agreement against AI-only assessment, devising human-led, AI-assisted models and advocating for intelligibility as the standard reference point instead of native speaker reflective competence. The study concludes that existing AI writing assessment tools will not be appropriate for high-stakes Indonesian ELT assessment without adaptation because they devalue local linguistic identities through systematic 'cultural erasure,' as one lecturer called it.

Keywords: algorithmic bias, automated writing evaluation, decolonial assessment, culturally situated writing, ELT assessment

Introduction

Artificial intelligence (AI) has recently been used in language assessment, and it is growing rapidly across the English Language Teaching (ELT) contexts around the world, including at many universities in Indonesia. More and more, tools like Grammarly and Turnitin are utilized for their efficiency, objectivity, and ability to give students instant feedback on their writing. Yet, the foundations of these automated systems have come under scrutiny through a body of scholarship. Studies have found that AI writing evaluators are mainly trained on standard native-speaker (US/UK) English corpora. This embeds linguistic norms as treating deviations therefrom as errors (Koltovskaia, 2020; Barrot, 2023). Such

algorithmic bias poses an issue for postcolonial ELT contexts, where varieties of English produced by English learners in these areas are legitimate products of local linguistic and cultural resources. This has led to a rise in decolonial approaches to assessment, questioning hegemony in academic quality synonymous with dominant English language standards and advocating for socially just assessments that value students' different voices (Waddington, 2024); Rosa & Flores, 2017). Compared to the Indonesian context, recent studies of compulsory English textbooks show positive trajectories toward decoloniality elements of representing the Indonesian population more culturally grounded and conceptualizing English as an intercommunicative medium between non-Western nations (Candraningrum, 2026). However, at the same time that these shifts have occurred, language assessment practices in general, and computer-automated ones in particular, have lagged behind (Wahyuningsih, 2024; Widodo, 2018) to continue punishing locally situated expressions exhibiting code-switching features / pragmatic transfer elements stemming from a knowledge of indigenous rhetorical patterns.

Awareness of AI bias in writing assessment continues to increase globally, however, the empirical database remains limited and fragmented, with few high-quality, internationally relevant studies employing a variety of English dialects / educational contexts. Past studies have recorded algorithmic bias against English language writing by Indian, Nigerian and Chinese English speakers (Markl, 2022; Giray, 2024), evidence that AI detection tools and automated scoring systems flagging / penalizing texts written in a non-English native tongue status are less likely to be flagged. For example, a study determined that grammar correction tools such as Grammarly snowballed AI detector false positive alerts, with an increase in non-native versus native English speakers at risk of being falsely flagged (Giray, 2024). Automated speaking and writing tests have also been reported to yield accurate scoring for non-dominant accented learners (Tajeddin et al., 2022) as well as linguistically diverse populations in general (Kaur et al., 2017). This fuels the reproduction. Yet, there are very few empirical studies that deal with Indonesian English as a one of varieties. However, no previous research has targeted as the primary subjects of AI bias studies English education students in university level inIndonesia. Further, most relevant research does not use student writing on local topics (regional traditions, indigenous values / community-based issues) but rather (a) isolated decontextualized sentences and/or (b) artificially-written texts (Abdul Rahman et al., 2023). Besides, there has been no previous research that has compared Grammarly and Turnitin scores to human lecturer scoring using a rubric which allows for culturally-located features such as code switching, pragmatic transfer / indigenous rhetorical patterns. This gap is needed for Indonesian decolonial ELT but increasingly, it emboldens the need of assessment practices that do not condemn / disregard local linguistic identities. Although there have been some textbook analyses that suggested the emergent decoloniality in Indonesian English language materials (Candraningrum, 2026), assessment practices, particularly those mediated by AI, remain under-explored.

This study is contextualized in the English Language Education Department at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar where students were directly prompted to combine local Balinese cultural values into their English output. The following research questions have been researched: To what extent do Grammarly and Turnitin scores penalize culturally situated features of English as a foreign language writing from Indonesian students pursuing degrees in English Language Education, and how do lecturers understand this bias? The study compares AI-generated scores from two writing support applications, Grammarly and Turnitin, against a baseline rated by human beings across sixteen student essays to answer this question. Existing literature on algorithmic bias in automated writing evaluation has mostly focused on the standardized and native English like scoring of second language production, which this study expands by providing evidence of AI bias against Indonesian English located locally. The results also offer implications for decolonial assessment theory, specifically by illustrating the ways in which AI tools developed on native-speaker norms create inequities for multilingual writers who use local linguistic resources. Moreover, this study provides contextualized recommendations for Indonesian ELT departments on how to responsibly use AI in writing assessment with reference to the rubric design, lecturer training and overcoming challenges posed by current AI tools as they relate to local English varieties.

Research Method

Design

This study used a sequential explanatory mixed-methods design with quantitative and qualitative data collection in two phases. It was measured first with the quantitative phase comparing Grammarly, Turnitin scores and a human-rated baseline that measured the extent of AI scoring bias. The qualitative phase subsequently clarified that bias by conducting semi-structured interviews with lecturers to explain what the form of that bias was and where it came from. This design was appropriate because the quantitative data by itself could not explain why bias occurred nor how lecturers interpreted culturally situated aspects, and the qualitative nature added meaning to statistical results.

Participants

Participants consisted of two groups. The first group was 16 students from the English Language Education Department at Universitas Hindu Negeri I Gusti Bagus Sugriwa Denpasar. They were then chosen mostly from semesters 5 to 7, which was a post-intermediate to advanced scientific writing class. One culturally situated writing sample, an argumentative essay on one of their local issues was submitted by each student. The second group consisted of three instructors from the same department. Purposive sampling of lecturers was based on 1) teaching writing / language assessment-related courses; 2) having a minimum two years' experience in assessing student writing, and having familiarity with both AI tools (Grammarly and Turnitin); and 3) being familiar with local Indonesian English varieties typical of Balinese learners.

Instruments

Three instruments were used. Student writing samples became the main data source. Each student wrote one 300-400 word essay based on a culturally situated topic submitted and approved by their lecturer. They aimed to generate locally grounded utterances. Secondly, the operationalization of AI scoring rubrics represented Grammarly's score (0-100) and Turnitin's Similarity Report plus grammar feedback recorded for each essay in numerical form. Third, a semi-structured interview guide for the lecturer participants. It examined their impressions of AI accuracy, examples of perceived bias, and their views on whether features that were culturally situated should be penalized / accepted.

Data Collection

Collection of data followed three steps in sequence. First step, this study collected sixteen anonymous essays from students, coded S1 through S16. It then submitted each essay to Grammarly (premium version, used consistently the same) and Turnitin only once. Both AI scores were then saved in a data matrix. Step two was that the three lecturer participants then each independently rated the same 16 essays, using a developed rubric that permitted culturally situated features. AI scores were hidden from lecturers. Each essay received three human scores that were averaged to see what humans thought about the score. The following two components were core to each of the three sections completed by each of these lecturers. Step three was that each lecturer engaged in a 30-45 minute semi-structured interview conducted using Indonesian respectively. Interviews were recorded with consent and transcribed verbatim. In this case, lecturers looked at some of the essays where AI and human scores differed by more than 10 points and wrote a short explanation for why they awarded that score.

Data Analysis

This study involved parallel quantitative and qualitative analysis processes. For each essay the distance between each AI score (Grammarly and Turnitin) and human baseline was calculated as a bias score. If the bias score was positive, that meant AI gave it more points than humans. If it was negative, that meant AI penalized the essay compared to humans. This produced descriptive statistics (mean, range, standard deviation) on bias scores for all 16 essays. In order to assess whether the mean bias score was statistically different from zero, thi study performed a paired t-test / Wilcoxon signed-rank test depending on normality. They were further disaggregated by essay topic and by individual linguistic features (code-switching, pragmatic markers) that yielded the largest penalties. Interview transcripts were thematically analyzed following the six-phase framework by (Braun & Clarke 2006). Using a hybrid coding approach, two coders coded the transcripts using both deductive codes derived from the

quantitative findings and inductive codes emerging from the transcript data. Inter-coder reliability was quantified using Cohen's Kappa and disagreements between codes were resolved through discussion. These themes were then combined with quantitative findings to respond to the research question concerning how widespread AI bias was (quantitative), and through qualitative data, also to inform whether / not lecturers saw it. Integration was achieved through a display table matching bias patterns with lecturer explanations.

RESULTS AND DISCUSSION

Results

Quantitative Findings

In this quantitative phase of the study, Grammarly and Turnitin scores were compared to human lecturer baseline scores across 16 culturally situated student essays (S01–S16). All written essays were rated separately by three lecturers with the holistic rubric (Instrument 3) and an average human score was derived for each essay. Bias scores (AI score-Human baseline score) were calculated, with negative values representing essays that received a higher penalty from AI compared to human raters. To standardize scores for comparability we converted them to 0-100 scale. Inter-rater reliability between the three lecturers was determined using Fleiss' kappa, resulting in $\kappa = 0.78$ (Web Table). Human baseline scores were between 50 and 92.5 (mean = 73.8, SD = 11.2). A reflective essay incorporating some untranslatable Balinese terms with clarity was rated the best (S08, score 92.5). The lowest rated essay (S14, score 50.0) featured various grammatical errors that independently obscured understandability, independent of its culturally situated properties.

Finding 1: Mean Bias Score for Grammarly

Table 1: Summary of Grammarly Bias and Flag Data Compared to Human Raters

Metric	Value
Mean bias score (Grammarly vs. humans)	−11.8 (SD = 7.9, range = −27 to +2)
Essays scored lower by Grammarly	15/16 (93.75%)
Essays scored higher by Grammarly	1/16 (6.25%)
Most negative bias (essay S03)	−27
Least negative bias (essay S09)	−3
Total flags generated by Grammarly	247
Flags due to culturally situated features	112 (45.3%)
- *Balinese/Indonesian words flagged as misspelled*	47
- *Politeness markers flagged as wordy*	31
- *Indirect rhetorical structures flagged as unclear*	22
- *Cultural terms flagged as needing definition*	12

Using a 0-100 scale and the default settings for Grammarly, essays received an average negative bias score from Grammarly of -11.8 (SD = 7.9, range = -27 to +2), which is equivalent to scoring students' essays almost 12 points lower than human lecturers on an essay scale from 0 to 100. That's an enormous penalty, nearly a full letter grade on the A-F scale. Looking at the bias scores for essays as individual pieces of work, 15/16 (93.75%) were scored lower (more biased) by Grammarly than by a human rater. I actually found only one (S12) that received a higher than human baseline score from Grammarly, albeit by only +2. S12 was an essay about Balinese tourism that involved a position and reasoning, and written in highly formulaic native-English, with few culturally situated features. This indicates that Grammarly punishes writing situated in a cultural context, but slightly favors / aligns with human raters when such features are scarce. The most negative bias for Grammarly was for essay S03 (bias = -27), which had the highest incidence of code-switching like Balinese terms 'melasti,' 'pecalang,' and 'odalan.' All three terms proved to be misspellings, as noted by Grammarly, and many pragmatically appropriate politeness markers were defined as 'awkward phrasing'. For Grammarly, the most minimal negative bias was found for essay S09 (bias = -3) which had an only one culturally situated feature ('Tri Hita Karana') but otherwise resembled more traditional academic English with direct thesis placement and little code-switching. In total, Grammarly generated 247 flags for the 16 essays she studied: 112 (45.3%) of them

were directly due to the culturally situated features, compared with standard grammatical errors such as missing periods. Of these 112 flags, 47 are Balinese or Indonesian words flagged as misspelled ('melasti' suggested to be a 'melastic,' an imaginary term) 31 were politeness markers flagged as wordy (for suggestion to become simply 'please'), 22 were indirect rhetorical structures flagged as unclear and twelve were cultural terms that had been flagged to require definition even though they have already been defined elsewhere in the essay.

Finding 2: Mean Bias Score for Turnitin

Table 2: Summary of Turnitin Bias and Flag Data Compared to Human Raters

Metric	Value
Mean bias score (Turnitin vs. humans)	−6.9 (SD = 5.8, range = −21 to +4)
Essays scored lower by Turnitin	14/16 (87.5%)
Essays scored higher by Turnitin	2/16 (12.5%)
Most negative bias (essay S14)	−21
Least negative bias (essay S02)	−2
Total flags generated by Turnitin	178
Flags due to culturally situated features	68 (38.2%)
- *Politeness markers flagged as too informal*	29
- *Indirect thesis placement flagged as unclear*	24
- *Specific cultural terms flagged as non-standard*	15
- *Generic language suggestions*	5
Note on Balinese terminology	Not flagged as spelling errors (unlike Grammarly)

Turnitin had a mean bias score of -6.9 (SD = 5.8, range = -21 to +4), meaning that Turnitin marked student essays on average 6.9 points lower than human lecturers did. Grammarly's negative bias was approximately 4.9 points more detrimental than that of Turnitin. This data shows either that culturally located features are marginally better tolerated by the scoring algorithm used by Turnitin / an interest in organization and plagiarism as opposed to grammar and vocabulary reflected in surface errors. Analysis of individual bias scores for essays from the original dataset showed that 87.5% (14 out of 16) received lower Turnitin score than ratings given by human evaluators. The top two essays (S04 and S11) had Turnitin scores roughly 3 and 4 higher than the human baseline. S04 was basically a reflective essay on Galungan that mentioned only one Balinese word ('penjor') and defined it straight away in parentheses. S11 was an indirect rhetorical structure values-based essay of Tri Hita Karana that had no indexical code-switching, Turnitin's feedback classified the introduction as 'unclear organization' but rated the overall score just a fraction above the Human baseline, likely because there were no spelling / grammar errors by traditional markers. Turnitin had its biggest negative bias for essay S14 (bias = -21), which contained a high proportion of pragmatic transfer features, such as 'please give me permission,' 'I ask for your blessing' and 'if you don't mind sir.' These were labelled as non-standard phrasing and potential formality issues (Turnitin), which the computer could characterise such polite requests as being too colloquial for academic writing. The lowest negative bias for Turnitin was displayed by essay S02 (bias = -2) which used a single Balinese term ('banten') with a parenthetical definition and adhered to formal academic English conventions otherwise. Turnitin assigned a total of 178 individual flags to the 16 essays, as seen in the following table (68 or 38.2% of items were attributed to culturally situated features). The first set of flags were 68 markers suggesting that the text was too informal, including 29 marks for politeness (please give me permission, flagged as too informal for academic writing) and 24 marks for indirect thesis placement (unclear thesis placement). Five were generic language ('consider stating your central argument earlier'), and 15 specific cultural terms flagged as non-standard ('try a more popular name'). Turnitin, unlike Grammarly, does not identify Balinese terminology as spelling errors. This can be a major distinction in how each tool handles non-standard vocabulary.

Finding 3: Statistical Significance of Bias Scores

Table 3: Statistical Significance of Bias Scores for Grammarly and Turnitin

Metric	Grammarly	Turnitin
Shapiro-Wilk test (normality)	W = 0.94, p = 0.32	W = 0.91, p = 0.18
Mean bias score	–11.6	–6.9
95% Confidence interval	[–15.9, –7.7]	[–9.99, –3.90]
Paired t-test vs. zero	t(15) = –5.97	*(implied significant)*
Statistical significance	p < 0.001 (99%+ certainty)	p < 0.001 (99%+ certainty)

Normality assumptions were checked before performing parametric tests. Data for bias scores of Grammarly (W = 0.94, p = 0.32) and Turnitin (W = 0.91, p = 0.18) passed the Shapiro-Wilk test confirming that the pairwise t tests were appropriate analyses to perform on this data set. The researchers compared the mean bias score for Grammarly against zero using a paired t-test, which showed that it was significantly different from 0 (t (15) = -5.97 p 99%+ certainty. Grammarly and Turnitin exhibited an average bias of -11.6 (95%CI: -15.9 to -7.7) and -6.9 (95%CI: -9.99 to -3.90), respectively, which we are therefore 95% confident that the true population mean bias lies somewhere within these ranges, assuming normality.

Finding 4: Specific Linguistic Features Most Penalized

Table 4: Specific Linguistic Features Most Penalized by Grammarly and Turnitin

Category / Feature	Grammarly Mean Bias (SD)	Turnitin Mean Bias (SD)	Additional Details
Code-switching essays (11 essays, 68.75%)	–15.2 (6.8)	–9.4 (5.1)	Most common terms are *melasti (7), pecalang (4), Tri Hita Karana (4), ngaben (3), odalan (3)*
No code-switching essays (5 essays)	–4.6 (3.9)	–2.3 (2.8)	Less variance than full group
Difference (code-switching vs. none)	Statistically significant: t(14) = 4.21, p < 0.01	Statistically significant: t(14) = 3.67, p < 0.01	Both two-tailed
Pragmatic transfer / politeness markers (8 essays, 50%)	–12.6 (5.9)	–7.8 (4.3)	Phrases flagged are *'please give me permission,' 'I ask for your blessing,' 'with all due respect,' 'with utmost humility'*
Indirect rhetorical patterns (6 essays, 37.5%)	–8.9 (4.2)	–5.3 (3.1)	Feedback is that thesis unclear, put main argument earlier
By essay prompt:			
- Prompt A (reflective, Balinese tradition)	–14.3	–8.1	Highest penalty; avg. 4.7 Balinese terms per essay
- Prompt B (argumentative, tourism & culture)	–9.2	–5.6	Moderate penalty
- Prompt C (values-based, *Tri Hita Karana*)	–10.1	–6.4	Moderate to strong penalty

Bias scores were disaggregated by specific linguistic features and three categories of culturally situated writing created the strongest AI penalties. Code-switching, or using Balinese / Indonesian words without translation, was the first and most heavily penalized category. Eleven essays (68.75%) contained code-switching. The mean bias score for Grammarly was -15.2 (SD = 6.8) compared to the Turnitin mean of -9.4 (SD = 5.1) over these essays. Mean bias for all code-switching essays for

Grammarly was -6.2 (SD = 3.0) while mean bias from Turnitin was -4.28 (SD = 1.16); whereas mean bias for the five essays with no code-switching only had a value of -4.6 (SD = 3.9) and -2.3 (SD = 2.8) respectively resulting in less variance overall between these scores alone per essay when compared to the full group of essays by rating determined using grammar checking tools. The increase in bias between the essays with code-switching and those without was statistically significant (two-tailed) for both Grammarly, $t(14) = 4.21$, $p < 0.01$, and Turnitin, $t(14) = 3.67$, $p < 0.01$ The most common Balinese terms among the penalized Balinese were *melasti* (7 flagged essays), *pecalang* (4 flagged essays), *ngaben* (3 flagged essays), *odalan* (3 flagged essays) and *Tri Hita Karana* (4 flagged essays). Grammarly's suggestion for *melasti* were mostly *'melastic'* (which does not exist) or 'elastic' (semantically different), and *pecalang* became ones with totally un-matching meanings.

The second category involved pragmatic transfer (particularly politeness markers) with the second most penalties. Eight essays (50%) included pragmatic transfer features, including formulaic request items like 'please give me permission,' 'I ask for your blessing,' and with the honors of all due respect' and 'with the utmost humility, I plead. The mean bias scores for these essays were -12.6 (SD = 5.9) using the Grammarly API and -7.8 (SD =4.3) using Turnitin, respectively. In four different essays Grammarly's feedback highlighted 'please give me permission' as wordy / redundant and suggested the student just omit 'please' / edit to read 'permit me.' The feedback indicated that similar phrases were 'informal' / 'non-academic' in three essays, where one read: 'This phrasing can be too informal for academic writing. You may want to consider a more straightforward wording. Significantly, all three lecturers rebutted these characterizations by explaining that in a Balinese Hindu context issuing commands without politeness markers (first names / terms of addressing elders) to someone older / in a position of authority can sound rude / disrespectful. Indonesian rhetorical patterns where introductions move slowly without directly stating the thesis statement, were the third most penalized category. These patterns were found in six essays (37·5%). Grammarly generated a mean bias score of -8.9 (SD = 4.2) for these essays, while Turnitin generated a mean bias score of -5.3 (SD = 3.1). This is often evident in Turnitin's feedback on these essays ('Thesis statement unclear,' / 'Consider putting your main argument earlier). Turnitin highlighted he first paragraph of one essay (S11): 'Your thesis statement comes very late in the introduction. Other academic readers start to expect to see your main argument from the first few sentences. Yet, indirect introductions were marked by the three lecturers as 'acceptable' and 'culturally appropriate in Balinese context': L2 writing: This student follows Indonesian academic conventions. Once you read the background, the thesis is pretty clear. No penalty.' Disaggregating bias scores by essay topic indicated that Prompt A (reflective essay on Balinese tradition) was penalized the most using either AI tool (Grammarly mean bias -14.3, Turnitin mean bias -8.1) as these essays contained the greatest amount of Balinese code-switching (average of 4.7 Balinese terms per essay). Moderate penalties were observed for Prompt B essays (essay for an argumentative essay on tourism and culture) (Grammarly mean bias -9.2, Turnitin mean bias -5.6); moderate to strong penalties were found for the Prompt C essays (values-based essay on *Tri Hita Karana*) (Grammarly mean bias -10.1, Turnitin mean bias -6.4).

Qualitative Findings

For the qualitative phase, semi-structured interviews were conducted with three of the lecturers (L1, L2 and L3). Interviews took place for 30-45 minutes and were conducted in Indonesian by the preference of the participants. Interviews were recorded, transcribed verbatim and translated into English for the purpose of analysis. Thematic analysis was conducted using the framework by Braun & Clarke (2006) with inter-coder reliability level of $\kappa = 0.84$ (indicating very good agreement) between two independent coders. Four themes were identified through the analysis.

Finding 5: Lecturers' General Perception of AI Accuracy

The three lecturers, however, provided a condemnation of AI write assessment tools. L1, a senior lecturer with twelve years of teaching experience said; 'I use Grammarly to catch basic spelling errors and comma errors. That is all. Its score is something I would never trust. It is so clueless about what my students are trying to say. Last semester a student came to me perplexed because her essay scored a 62 in Grammarly. I read the essay. The structure was smooth, the arguments were straightforward and she

used very good Balinese terms with explanations. I gave her an 85.' As reported, L2, an expert on language assessment, and one of its contributors to a recent publication about automated essay evaluation, told: 'Turnitin is great for checking plagiarism. For writing quality? No. Absolutely not. It has been able to help in flagging errors in sentences but I have witnessed it mark correct sentences wrong just because they do not follow American conventions. The issue is that it train on American academic writing It has no idea there are English varieties in Indonesia. My students are writing for neither an American audience. And they are writing for me as Indonesian lecturer.' L3, the youngest at five years' experience, took a more positive view but still highlighted restrictions: L3 'I think AI can be useful as a draft checker. For example, a student can run their essay through Grammarly before submitting it to me. They can dig into its suggestions around grammar, like subject-verb agreement and comma splices. Context, culture and intention can only be judged by a human.' In response to the open question of how often they use AI tools in their own assessment practices, L1 said 'rarely, maybe once a month' (with respect to Grammarly), L2 said 'for every final paper but only for plagiarism' (with respect to Turnitin), and L3 said 'occasionally to check my own writing before submitting to journals but never for grading students' (with regard to both). All three involved in the session were unanimous that AI tools are not appropriate for high-stakes assessment in Indonesian ELT contexts, especially when it comes to culturally situated writing.

Finding 6: Specific Examples of Perceived Bias from Lecturers

All three lecturers spotted examples of bias and wrote extensive commentary when looking at essays with a divergence greater than 10 points between the marks given to AI-generated and human-graded content. Regarding the essay S03, L1 gave lengthy comments at -27 on Grammarly (Grammarly score 58, human mean 85) and -11 on Turnitin's bias (-11; Turnitin Score 74, human mean=85). 'This student wrote the melasti ceremony. Well, she had five iterations of the word 'melasti' weave into her work. Another term she used was 'pecalang' (which are traditional security officers in Bali) and 'odalan' (temple festival). This list leaves something to be desired: Every single one of these words was flagged as a spelling error by Grammarly. But these are not errors. These are Balinese words. This student is writing about her own culture. What would she gain by translating melasti into English? There is no precise translation in English. In the next sentence, she explained what it means: 'Melasti is a purification ceremony before Nyepi. That is writing not to be dissed. It is unnecessary for Grammarly to suggest a change from melasti to melastic. There is no word 'melastic' in any dictionary. Had the student accepted that recommendation, however, she would have made a real mistake in her essay.' L2 continued with essay S07 (Grammarly score 58, human average 78; Grammarly bias -20; Turnitin score 70; human average 78; Turnitin bias -8) which wrote 'before Nyepi, we do melasti ceremony at the beach to purify the sacred objects'. Another common suggestion from grammar-checker Grammarly is to replace 'melasti' with 'melastic' a term that has nothing to do with Balinese culture. It is harmful because a student who does not know can take these suggestions seriously. I have seen this happen. One student writes 'upacara' in her essay, Grammarly says to change that to 'upgrade,' and she does because the AI knows better. Look, Grammarly knew more than she did. I asked her why she changed it, and she said, 'Grammarly says this was incorrect. Such an AI-based tool tells a student 'your culture is spelled wrong,' and therefore they became less confident in their own cultural knowledge.'

L3 explained this for essay S14, which had a Grammarly bias of -18 (Grammarly score 55, human average 73) and a Turnitin bias of -15 (Turnitin bias of -8, Turnitin score 58), traditional paper average low credit student management strategy. 'These were flagged as wordy by Grammarly who suggested: 'because please and I ask your blessing' to be omitted. Turnitin flagged them as informal writing. However in Balinese Hindu, courtesy is necessary. These polite forms are used when we talk with someone older or in the position of privilege. She was being respectful. This is a pattern I have been observing for months. Compared with American students, my students use 'please' more frequently. American English is more direct: 'Give me permission.' Indirectness is preferred in Indonesian English: 'Please allow me to.' Both are correct English. They are just different varieties. AI, which was not trained on Indonesian English, does not understand this. It was trained by American and British English.' In essay S11, with a bias of -4 based on Grammarly (score 76, average human 80) and +3 for Turnitin (score 83, average human 80), L2 comments: 'This essay starts with a three paragraph background before

offering up the thesis. Turnitin labelled this as 'unclear thesis on placement' and suggested placing the thesis in the first paragraph. But this is how the students from Indonesia learn to write. We provide context first. We build the argument gradually. This should be followed by the thesis once the background is complete. Which is unlike American academic writing. But different does not equal wrong. For international journals, I advise my students to use a direct thesis statement. But for my class, for writing about Balinese culture, indirect structure is perfectly acceptable.'

Finding 7: Lecturers' Stance on Penalizing Culturally Situated Features

All three lecturers agreed, however, without exception that culturally situated features should not be penalized, they provided different boundary conditions and rationales. 'I have only one question when judging a culturally situated feature,' L1 said: 'Do I understand what the student is trying to convey? If yes, no penalty. If not, we should have a problem. But 'melasti'? I understand perfectly. 'Please give me permission'? I understand perfectly. An indirect introduction? I understand perfectly. So no penalty. The student has communicated successfully. That is the goal of language.' L2 added: 'Not conformity to American norms. Not conformity to Grammarly's preferences. Not compliance with Turnitin's definition of 'clear organization.' Could I as a fairly decent and fluent English speaker and familiar with Balinese culture even get the message? If your answer is yes, the student has won. If the answer is no, then we need to talk about how it all went wrong. But, from my perspective, if I am struggling to read something a student wrote it is because they code-switched / used politeness markers not most of the time. It is often because of actual grammatical mistakes: tense mistakes, subject-verb mismatch and omission of articles. The problem with AI is that it does not know local diversity, hence cannot differentiate these categories. It treats both as errors. Ok, humans can and have to do this.'

L3 did the most subtle here: 'There are really two very different things we need to distinguish between. First, local variety 'We did ngaben ceremony' (a local Balinese term but this has to be explained xplicitly), 'I beg for your blessing' (politeness marker appropriate to Indonesian culture), before discussing the thesis let me introduce these in context. These are true characteristics of Indonesian English. Secondly, true mistakes, I go to temple yesterday (tense) She do not understand (subject-verb agreement), Melasti is a ceremony that purifies (missing article before 'ceremony'; it should be 'a ceremony'). This is where AI fails, as there is no understanding of what a local variety is. It basically considered both categories as an error. This difference needs to be drawn by humans. For example, if a student uses the term 'melasti' without context, I will ask them to define it. But I would not take points away. If a student writes melastic because Grammarly suggested it, I would mark the paper down because that is not a word. My students are learning the wrong English because of AI. This is the complete opposite of how assessment should act. All of the three lecturers concurred that intelligibility to an Indonesian English lecturer should be the main test for where to draw the line with localised variety versus unacceptable error. Just to summarize, L2, 'Not to speak like an American / a British person. We need to communicate clearly. A text is a success if an Indonesian English speaker read it and knows what that text means. AI is ignorant of this because AI was constructed on a native-speaker-model.'

Finding 8: Lecturers' Recommendations for AI Use in Indonesian ELT Assessment

In particular, all three lecturers shared detailed recommendations on the role of AI tools in Indonesian ELT assessment. The first recommendation from all three lecturers is to apply a combination of AI and human. L1 said, 'Artificial intelligence can be a second pair of eyes. A student can run a check of his draft using Grammarly. The lecturer compares the work with a student who has already been checked and Turnitin returns to them the original file. But the thing is, the final score has to be given by a human. Artificial Intelligence should never replace the lecturer. It should assist the lecturer. If my department required me to use Grammarly scores as a grading method, I would not comply. That would be denying my students the fair chance, again in particular for those who write on Balinese culture.' L2 continued, 'The optimal model is human-led, AI-assisted. The human lecturer reads the essay, forms a conclusion and uses AI only in cases of bullets for missing commas / replicating files. Explain that the AI should not ever generate score. The human generates the score. The AI is just a tool.' And the second one, from L2 and L3, was successful adaptation of AI tools to local varieties of English.' My guess for L2

was, 'Add Indonesian English to Grammarly' They have UK English, US English, Canadian English, Australian English. So they need to include Asian Englishes as well. Indonesian English, Malaysian English. Philippine English. They are real variations and genuine patterns. By the way, if Grammarly have Indonesian English setting, 'please permission me' would not be flagged as wordy. It would recognize this as an actual politeness strategy. For example, it would not recognize 'melasti' as a typo. It even identifies it as a Balinese cultural word. This is technically possible. This is already something Grammarly has in place for different varieties of the them. They just have to keep adding more types.' L3 concluded: 'AI tools should be used with utmost care until adapted. I have always told my students 'Grammarly can help with typos, and other basic grammar issues. However it can recommend you vocabulary / style, but do not accept any please until you ask me. Grammarly does not even understand Balinese culture.'

The third, presented by L1 and L3. L1 told, 'We should have a written policy in our department that says very plainly, we are not going to be penalizing culturally situated expression in AI assisted assessment. Inform students that they are able to use Balinese terms and explain their meaning as part of the performance. Students must be informed that politeness markers are not mistakes. The students are to be informed that indirect rhetorical structures can also be used for writing in Balinese topics. This should be given to students at the start of any writing course. The rule should seize on the opposite drawback, college students relying too closely on what an AI suggests, thus destroying how they write,' added L3. We need to teach AI literacy. Our students need to be informed that AI is far from perfect. Students must be aware of the fact that AI is not familiar with their culture. Students need to realize that their knowledge of Balinese pipelines has its worth and should not be wiped off by algorithm trained on American English. L2 wrapped things up with a parting shot, 'The future is not ambiguous if we ignore it. Indonesian English will be continue getting penalties from AI tools. Students will try to stop writing with their culture. They will know that 'melasti' is a mistake. They will not find out that everything polite is a lot of words. They will be taught that they are wrong about how to organize ideas. This is cultural erasure. We fight back against it, in our classrooms, and without department policy.'

Discussion

Findings from this mixed-methods study provide evidence that AI writing assessment tools like Grammarly and Turnitin unfairly penalize EFL, both in terms of systematically evaluating culturally situated features of Indonesian EFL writing. How dramatic this kind of negative impact can be, once again as shows by the average score penalties across the 10 best-performing local varieties (mean penalty of -11·8 points from Grammarly affecting 93.75% of essays and mean penalty of -6.9 points from Turnitin affecting 87.5%) for hypotheses are affected in a systematic way and not just through an occasional one-off case or mistake. These findings confirm and build on existing literature showing algorithmic discrimination against World Englishes including Indian, Singaporean / Nigerian English (Markl, 2022; (Giray, 2024; Rosa & Flores, 2017). The fact that even Turnitin's plagiarism-focused algorithm, which one assume would be less susceptible to linguistic bias, still lent negative penalties, suggests that native-speaker norms are pervasive across different strategies of AI writing (Koltovskaia, 2020; Barrot, 2023). Details on the differing penalty patterns of Grammarly and Turnitin indicate aspects of how each tool defines 'correctness.' The corrective bias of Grammarly (-11.8 versus -6.9) is more severe, because its main app is a grammar and style checker which has been trained predominantly on American- and British English Githubs, a limitation described in the automated writing evaluation literature (Dikli, 2006; Weigle, 2013). As one lecturer pointed out, the algorithm's mindless replacement of a major Balinese cultural concept, 'melasti,' with gibberish virgin name 'melastic' is an example of how supposedly helpful corrections can do damage to student writing by mistakenly inserting errors into their prose. This matches with recent findings reporting that AI tools frequently label appropriate characteristics of non-native varieties of English as errors in need of correction (Markl, 2022; (Tajeddin et al., 2022). The comparatively milder bias we found with Turnitin (Manley, 2023) owe to its preoccupation being plagiarism detection and more generalized text-matching (rather than grammatical prescriptivism), but it still tends to flag Indonesian politeness markers as 'too informal' for academic writing, showing that even tools not aimed at grammar checks can internalize and amplify Anglo-American academic conventions (Giray, 2024).

The ordering of penalties associated with linguistic features also reveals more insights into the aspects of culturally situated writing that appear the most susceptible to algorithmic bias, with code-switching receiving the most severe estimated penalties (Grammarly -15.2), followed by politeness markers (-12.6) and indirect rhetorical structures (-8.9). The implication of this finding is quite important in Indonesian ELT contexts where, as the lecturers explicated, indirectness and hierarchical politeness markers are not stylistic choices, they are culturally required aspects of proper communication (Widodo, 2018; Wahyuningsih, 2024). That, when Grammarly warns a student that 'please give me permission' is wordy / Turnitin recognizes similar expressions as too informal, they are not identifying an objective error in the paper but rather imposing cultural expectations about appropriate academic register. This pressure embodies what critical applied linguists have referred to as the ongoing dominance of inner-circle English standards in assessments (Waddington, 2024; Rosa & Flores, 2017). The 30 EFL lecturers' overwhelming rejection of an assessment approach in which only AI substitutes the classroom teacher, and their support for using AI to assist human-delivered teaching, is consistent with recent research that highlights the need for appropriate models of integration of AI tools into Indonesian EFL contexts. Recent studies have shown that AI tools are much better at technical aspects like grammar and mechanics, but a weaker correlation on holistic dimensions such as coherence and organization of content (Koltovskaia, 2020; Barrot, 2023). This backs the view of the lecturers that AI can be a useful 'second pair of eyes' for detection of surface-level error but cannot make judgments on meaning, cultural appropriateness, and rhetorical effectiveness (Abdul Rahman et al., 2023). This insistence by the lecturers that intelligibility rather than native-speaker conformity is the standard for their evaluation, represents a decolonial approach to assessment, challenging the pervasive assumption that linguistic difference equals to deficit (Waddington, 2024).

The quotes of the lecturers regarding the conception of prejudicial pedagogical effects that algorithmic bias can yield is to be taken seriously. For example, the record of a student changing 'upacara' to 'upgrade' because Grammarly erroneously suggested it highlights how AI tools can actively undermine students' confidence in their cultural and linguistic knowledge. Considering previous research, revealing that EFL students had a tendency to use AI tools for superficial mistakes without an adequate understanding of their shortcomings (Giray, 2024), this finding becomes particularly worrying. When AI tools repeatedly flag culturally appropriate expressions as mistakes, they communicate a message that students' local forms of talk are not legitimate in scholarly settings. One of my lecturers cautioned us that this is 'cultural erasure' which could result in students folding up and no longer expressing themselves but simply repeating formulaic native-speaker utterances (Rosa & Flores, 2017). The suggestions the lecturers provided for responsible integration of AI in Indonesian ELT assessment indicate potential solutions that combine technological efficiency with pedagogically and culturally appropriate approaches. Recent research implies that the proposed hybrid framework in which AI completes some initial technical screening but human judgment is maintained for final decisions would retain assessment quality while also allowing many of the efficiencies afforded by AI (Wahyuningsih, 2024). Tackling the governance gap revealed in studies of Indonesian lecturers' perspectives on AI integration (Widodo, 2018), the call for institutional policies specifically safeguarding culturally situated expression from punitive action based on AI output is accepted. Teaching students to be aware of the weaknesses of AI and judging AI's suggestions instead of accepting them blindly, as suggested by research that many Indonesian EFL students assessed before using AI tools, implies a deficiency in awareness on codes of bias (Candraningrum, 2026).

The theoretical relevance of this study encompasses decolonial assessment theory beyond Indonesian ELT. This study shows algorithmic systems in action, bringing into empirical focus theoretical assertions that World Englishes continue to be subordinate even as neutral technological systems rise by showing how linguistic hegemony is operationalized and reinforced through machine learning. The fact that human judgment by active lecturers systematically gave different scores than AI on features of writing attached to particular cultural traditions supports calls for assessment practices centered in local voices and standards (Tajeddin et al., 2022; (Kaur et al., 2017). This study offers evidence-based recommendations for Indonesian ELT policymakers and practitioners to develop institutional policies that protect cultural heritage via a contextualised, culturally responsible approach to technological innovation.

Conclusion

The current mixed-methods study presents evidence that both Grammarly and Turnitin are subjecting culturally situated features of Indonesian EFL student writing to statistically verified algorithmic bias resulting in systematic penalties against non-native English varieties diverging from inner-circle (US/UK) linguistic norms. The quantitative results provide evidence for a significant negative bias on most essays, with Grammarly imposing an average -11.8 point bias overall (93.75% of essays) and Turnitin imposing a -6.9 point penalty on 87.5% of essays (both $p < 0.001$). Code-switching also emerged as the most heavily penalised feature (Grammarly bias -15.2), followed by politeness markers (-12.6) and indirect rhetorical structures (-8.9), revealing that AI tools are treating culturally appropriate Indonesian-Balinese communicative strategies as objective errors. In the qualitative findings from lecturer interviews, all three of the lecturers uniformly dismiss models of assessment that only enact AI and agree to intelligibility to a context-aware Indonesian English reader vs native-speaker expectation. Lecturers described how penalising on elements such as Balinese lexical insertions, hierarchical politeness markers, and indirect thesis placement is a form of cultural erasure which undermines the linguistic identity of students leading to self-censorship of local knowledge. The study concludes that current AI writing assessment means would require significant adaptation to be appropriate for high-stakes Indonesian ELT assessment, as locally situated expressions are systematically devalued by the algorithmic architectures of commercially driven AI systems trained only on American and British English corpora. Accordingly, this study suggests a human-centered hybrid model with an AI-assisted system where artificial intelligence performs the function of technical 'second pair of eyes,' limited to off-the-top error detection while human lecturers make final scoring and judgements for cultural appropriateness drawing on their local linguistic and cultural expertise. Moreover, the results warrant institutional policies clearly banning punishments for culturally situated qualities, as well as AI literacy training helps Indonesian EFL students to critically assess algorithmic recommendations instead of swallowing them whole. In the absence of these adaptations, when unadapted AI tools continue to be used in Indonesian ELT assessment, linguistic colonialism will persist, as students are taught that their own culturally-bound ways of knowing and communicating are inherently wrong.

References

Abdul Rahman, N. A., Zulkornain, L. H., Che Mat, A., & Kustati, M. (2023). Assessing Writing Abilities using AI-Powered Writing Evaluations. Journal of ASIAN Behavioural Studies, 8(24). https://doi.org/10.21834/jabs.v8i24.420

Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy. Computer Assisted Language Learning, 36(4). https://doi.org/10.1080/09588221.2021.1936071

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

Candraningrum, D. (2026). Decolonial Fiction Pedagogy in ELT: Reimagining Memory, History, and Identity Through Laksmi Pamuntjak's The Question of Red. International Journal of Interdisciplinary Educational Studies, 21(1). https://doi.org/10.18848/2327-011X/CGP/v21i01/49-67

Dikli, S. (2006). An overview of automated scoring of essays. In Journal of Technology, Learning, and Assessment (Vol. 5, Number 1).

Giray, L. (2024). The Problem with False Positives: AI Detection Unfairly Accuses Scholars of AI Plagiarism. In Serials Librarian (Vol. 85, Numbers 5–6). https://doi.org/10.1080/0361526X.2024.2433256

Kaur, A., Noman, M., & Nordin, H. (2017). Inclusive assessment for linguistically diverse learners in higher education. Assessment and Evaluation in Higher Education, 42(5). https://doi.org/10.1080/02602938.2016.1187250

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. Assessing Writing, 44. https://doi.org/10.1016/j.asw.2020.100450

Manley, S. (2023). The use of text-matching software's similarity scores. Accountability in Research, 30(4). https://doi.org/10.1080/08989621.2021.1986018

Markl, N. (2022). Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. ACM International Conference Proceeding Series. https://doi.org/10.1145/3531146.3533117

Rosa, J., & Flores, N. (2017). Unsettling race and language: Toward a raciolinguistic perspective. Language in Society, 46(5). https://doi.org/10.1017/S0047404517000562

Tajeddin, Z., Saeedi, Z., & Panahzadeh, V. (2022). English Language Teachers' Perceived Classroom Assessment Knowledge and Practice: Developing and Validating a Scale. Profile: Issues in Teachers' Professional Development, 24(2). https://doi.org/10.15446/profile.v24n2.90518

Waddington, J. (2024). Questioning the native speaker construct in teacher education: Enabling multilingual identities and decolonial language pedagogies. In Questioning the Native Speaker Construct in Teacher Education: Enabling Multilingual Identities and Decolonial Language Pedagogies. https://doi.org/10.4324/9781003188896

Wahyuningsih, S. (2024). Does Artificial Intelligence (AI) Play Roles in Enhancing Academic Writing? Unravelling Lecturers' Voices in Indonesian Higher Education. Jurnal Pendidikan Progresif, 14(1). https://doi.org/10.23960/jpp.v14.i1.202436

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In Handbook of Automated Essay Evaluation: Current Applications and New Directions. https://doi.org/10.4324/9780203122761

Widodo, H. P. (2018). A Critical Micro-semiotic Analysis of Values Depicted in the Indonesian Ministry of National Education-Endorsed Secondary School English Textbook. In English Language Education (Vol. 9). https://doi.org/10.1007/978-3-319-63677-1_8