Multidisciplinary Journal of Science Technology Education and Management
Journal homepage: http://www.mjstem.org

Writing Under Suspicion: AI-Detection Anxiety, Writing Self-Censorship, and Humaniser Intention among Vietnamese EFL Learners

Quoc Khanh Nguyen[a], Nhat Quang Vo[b,*]

[a]The University of Da Nang, Da Nang City, Vietnam;
[b]Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam.

[*]Correspondence: Nhat Quang Vo (020401250009@st.buh.edu.vn).

Abstract

AI-writing detectors are increasingly used in educational assessment, but their learner-facing consequences remain insufficiently understood, especially for writers using English as a foreign language. This study examined AI-detection anxiety, writing self-censorship, and humaniser intention among Vietnamese EFL learners, compared learners reporting previous AI-detector flagging with those reporting no previous flagging, and tested the association between reported flagging experience and categorical humaniser intention. A cross-sectional survey was completed by 142 adult learners and analysed using descriptive and comparative methods. Data were analysed in R using frequencies, percentages, means, standard deviations, Cronbach's alpha, Welch's independent-samples t-tests, Cohen's d, chi-square analysis, and Cramér's V. AI-detection anxiety was the most strongly endorsed construct (M = 3.65, SD = 0.98), followed by humaniser intention (M = 3.13, SD = 1.01) and writing self-censorship (M = 2.97, SD = 0.98). Learners reporting previous flagging reported higher anxiety than those reporting no previous flagging, $t(87.52) = 2.22$, $p = .029$, $d = 0.41$. No significant group difference emerged for self-censorship, and reported flagging experience was not significantly associated with categorical humaniser intention. The study distinguishes an emotional response, a restrictive writing response, and a prospective tool-use intention rather than treating them as equivalent evidence of misconduct. The findings indicate substantial detector-related anxiety without showing that such anxiety uniformly co-occurs with writing self-censorship or humaniser intention. The findings support caution in treating detector scores as conclusive evidence and suggest the importance of transparent authorship-review procedures.

Keywords: AI detection, EFL writing, academic integrity, writing self-censorship, humaniser intention

1. Introduction

Generative artificial intelligence has rapidly entered second-language writing, supporting brainstorming, organisation, correction, and revision. Recent research involving Vietnamese

Received July 12, 2026 Revised August 20, 2026 Accepted September 02, 2026
EISSN: 3121-9292
ISSN: 3121-6552

EFL learners has examined the use of AI writing tools and learners' perceived benefits and challenges, including support for idea development, language editing, and writing efficiency alongside concerns about overreliance, plagiarism, and writing authenticity (Bui & Tong, 2025; Meniado et al., 2024; Thi et al., 2025). Teacher-focused research in Vietnam has likewise examined GenAI-related academic-integrity concerns and pedagogical responses (Cong-Lem et al., 2024). These developments complicate distinctions among legitimate assistance, independent authorship, and unacceptable text generation.

Institutions have responded through AI-use declarations, revised assessment policies, and automated detectors. Yet university guidance remains uneven, and detector outputs are probabilistic rather than direct proof of authorship (Moorhouse et al., 2023). Evaluations show substantial variation across tools and weaker performance after textual modification (Perkins et al., 2024; Weber-Wulff et al., 2023). Liang et al. (2023) further found disproportionate classification of non-native English writing. Detector use therefore raises issues of accuracy, fairness, transparency, and learner trust.

Consequences may arise before any misconduct decision. Research on text-matching software shows that students can fear plagiarism allegations despite no intentional plagiarism and adopt counterproductive writing or citation practices (Goddiksen et al., 2024). Because text-matching systems and AI-writing detectors operate differently, this literature is used here as related evidence about learner responses to automated textual scrutiny, not as direct evidence about AI-detector performance or effects. Under AI-detection scrutiny, learners may similarly worry about evaluation and alter aspects of their writing. These responses must remain distinct: anxiety does not imply AI use, self-censorship is not ordinary revision, and humaniser intention is not actual use.

Despite growing research on detector accuracy, AI-assisted writing, policy, and academic integrity, limited quantitative evidence addresses learners' emotional and writing-process responses specifically to AI-detector scrutiny. To the authors' knowledge, existing research in Vietnam has not jointly examined AI-detection anxiety, writing self-censorship, and humaniser intention among EFL learners. This study addresses that gap through a learner-centred cross-sectional survey with descriptive and comparative analyses that distinguishes the three responses and examines reported previous AI-detector flagging without assuming detector accuracy or causality. Throughout this article, an AI-detection score refers to a detector's probabilistic output, whereas reported AI-detector flagging refers to a participant's self-report that they had been told, shown, or otherwise informed that their writing was classified as partly or fully AI-generated. Suspicion and formal accusation refer to human or institutional judgements rather than detector outputs, and confirmed AI use was not established in this study. These terms are therefore not treated as interchangeable. A false-positive classification would require evidence that a detector classification was incorrect relative to the text's actual provenance. Because detector outputs were not independently verified, this study does not identify or estimate false-positive cases. References to "false AI detection" or "false flagging" in the administered questionnaire therefore represent participants' feared or perceived misclassification, not researcher-verified false positives or formal accusations.

The study addressed four research questions:

1. What levels of AI-detection anxiety, writing self-censorship, and humaniser intention are reported by Vietnamese EFL learners?

2. Which AI-detection-related concerns are most commonly reported?

3. Do learners reporting previous AI-detector flagging report higher AI-detection anxiety and writing self-censorship than learners reporting no previous flagging?

4. Is reported previous AI-detector flagging associated with categorical humaniser intention?

2. Conceptual and Theoretical Framework

The framework integrates control-value theory with the chilling-effects perspective while maintaining a clear distinction among emotional, writing-process, and prospective behavioural responses. These perspectives are used as interpretive lenses to conceptualise and explain the observed patterns, not as a causal model linking reported flagging, anxiety, self-censorship, and humaniser intention.

2.1. AI-Detection Anxiety and Control-Value Theory

AI-detection anxiety refers to situation-specific apprehension that one's writing may be classified as AI-generated and subsequently questioned or penalised. It differs from general anxiety about artificial intelligence and from guilt associated with actual unauthorised use. A learner may experience it even when a text was independently written or AI assistance complied with institutional rules.

Control-value theory explains academic emotions through perceived control and value. Anxiety is plausible when an important evaluative outcome seems difficult to control (Pekrun, 2006, 2024). Writing affects grades and reputation, whereas detector logic and appeals may be opaque. Inconsistent performance and bias against non-native writers make concern credible without proving that any individual flag was wrong (Liang et al., 2023; Perkins et al., 2024; Weber-Wulff et al., 2023). A participant-reported flagging experience is therefore treated as an evaluative experience, not as confirmation of AI use.

2.2. Writing Self-Censorship and Chilling Effects

Writing self-censorship is the deliberate restriction or alteration of otherwise legitimate language because it is perceived as likely to attract suspicion. It differs from ordinary revision, which aims to improve accuracy, coherence, audience appropriateness, or clarity. Here, the primary goal is avoiding an anticipated detector judgement, potentially at the expense of lexical development, stylistic choice, or academic voice.

The chilling-effects perspective proposes that perceived monitoring and possible sanctions can inhibit legitimate communication (Büchi et al., 2022). AI detection is algorithmic scrutiny

because textual features become probabilistic authorship judgements. Students concerned about plagiarism allegations have prioritised avoiding software-detected overlap over sound writing and citation (Goddiksen et al., 2024). This text-matching evidence is treated as an analogue for how automated textual scrutiny may influence writing behaviour, not as evidence that AI detectors produce identical responses.

2.3. Humaniser Intention and the Integrated Framework

In this study, a humaniser refers to a rewriting or paraphrasing tool marketed or used to make text appear more human-authored to an AI-writing detector. The term is therefore restricted to this detector-management purpose rather than ordinary language editing or paraphrasing in general. Humaniser intention denotes stated willingness to consider using such a tool under detector-related concerns. It is a prospective intention, not observed tool use. Its ethical meaning depends on the provenance of the text, the function of the tool, disclosure requirements, and institutional policy.

Paraphrasing and simple textual manipulation can reduce detector accuracy (Perkins et al., 2024; Weber-Wulff et al., 2023), but technical studies do not show how often learners contemplate such strategies in response to feared detector misclassification. The framework therefore distinguishes anxiety as an emotional response, self-censorship as a restrictive writing response, and humaniser intention as a prospective behavioural intention.

3. Methodology

A quantitative survey design was used to describe the three focal constructs and compare learners according to previous self-reported flagging experience. The procedures and analyses were restricted to those needed to answer the four research questions.

3.1. Research Design and Setting

The study used a cross-sectional survey design with descriptive and comparative analyses. A convenience sample was recruited from English-related undergraduate and postgraduate programmes at one university in Ho Chi Minh City, Vietnam. The institution remains unidentified. The design supported description and group comparison, not causal inference.

3.2. Participants

The final sample comprised 142 Vietnamese EFL learners, all aged 18 years or older. Most were undergraduates (n = 131, 92.3%), and 11 (7.7%) were postgraduate students. Women represented 55.6% of the sample and men 38.7%; the remaining participants selected another identity or preferred not to answer. Most participants were aged 18–23 years (n = 121, 85.2%), and B2 was the most common self-reported proficiency level (n = 66, 46.5%). Participant characteristics are detailed in Table 1.

3.3. Instrument

The questionnaire included demographic and detector-experience questions and 15 five-point Likert items developed specifically for this study. Rather than being directly adapted from a single established scale, the items were informed by the theoretical and empirical literature underlying the three constructs. Five items assessed AI-detection anxiety, drawing on control-value theory and evidence concerning uncertainty and perceived risk under automated evaluation (Pekrun, 2006, 2024; Goddiksen et al., 2024). Five items assessed writing self-censorship, informed by the chilling-effects perspective and evidence of restrictive or counterproductive responses to automated text scrutiny (Büchi et al., 2022; Goddiksen et al., 2024). Five items assessed humaniser intention, informed by evidence that textual modification can affect AI-detector performance (Perkins et al., 2024; Weber-Wulff et al., 2023). In the survey, a humaniser was defined as a rewriting or paraphrasing tool marketed or used to make text appear more human-authored to an AI-writing detector. The humaniser items were framed around the perceived risk of detector misclassification and therefore captured stated willingness under hypothetical detector-related scenarios rather than reported or verified tool use. Responses ranged from 1 (strongly disagree) to 5 (strongly agree), with higher means indicating stronger endorsement. No formal expert review or pilot testing was conducted before administration. The questionnaire was administered entirely in English, so no translation or back-translation was undertaken. A separate item asked whether participants would consider a humaniser if self-written English might be falsely flagged (yes, no, or not sure). Previous self-reported AI-detector flagging experience used the same response categories and served as the grouping variable.

3.4. Data Collection and Ethical Considerations

Eligible learners completed an anonymous, self-administered survey after receiving information about the study and providing informed consent. Participation was voluntary, and respondents could discontinue before submitting the questionnaire. No names, student numbers, institutional email addresses, instructor names, or class codes were collected. Formal ethics committee review was not obtained before data collection. The study was not formally determined to be exempt from ethics review, and no separate institutional permission for recruitment or data collection was obtained. We therefore do not claim that ethics review was unnecessary under an institutional policy. The adult-only, anonymous, non-interventional design and informed-consent procedures are reported as participant safeguards, not as grounds for exemption from prospective ethics review. Data were analysed and reported only in aggregate form. The absence of prospective ethics review and institutional permission is acknowledged as a procedural limitation.

3.5. Data Preparation and Statistical Analysis

Analyses were conducted in R through RStudio. Screening covered duplicates, coding, consistency, and missingness. All 142 identifiers were unique and Likert values were valid. Six items each had one missing response. Item-level statistics used all available responses for each item. Construct scores were calculated as the mean of available item responses when at least

four of the five items were completed; this criterion yielded construct scores for all 142 participants. Given the minimal item-level missingness, Cronbach's alpha was calculated from pairwise available covariances so that each item pair used all respondents with valid responses to both items rather than excluding an entire case because of a single missing item. Frequencies, percentages, means, standard deviations, and combined agreement addressed RQ1–RQ2. For RQ3, 18 participants reporting uncertainty about previous flagging were excluded, leaving 83 reporting no previous flagging and 41 reporting previous flagging. Welch's *t*-tests and Cohen's *d* assessed group differences. Pearson's chi-square and Cramér's *V* addressed RQ4. Tests were two-tailed at $\alpha = .05$, with effect sizes considered alongside significance.

4. Results

The demographic and categorical variables contained no missing data. Six scale items, ADA3, ADA5, WSC2, WSC4, HI3, and HI5 each contained one missing response; therefore, their item-level statistics were based on 141 cases. All other item statistics used 142 cases. Table 1 shows that approximately three in ten participants reported previous flagging, while nearly two-thirds had personally checked their writing with an AI detector at least once.

Table 1: Participant Characteristics and Previous AI-Detector Experience

Variable	Category	n	%
Gender	Woman	79	55.6
	Man	55	38.7
	Another identity	3	2.1
	Prefer not to answer	5	3.5
Age group	18–20	67	47.2
	21–23	54	38.0
	24–29	18	12.7
	30 or above	3	2.1
Educational status	First-year undergraduate	31	21.8
	Second-year undergraduate	31	21.8
	Third-year undergraduate	35	24.6
	Fourth-year or later undergraduate	34	23.9
	Postgraduate student	11	7.7
Self-reported English proficiency	A2	18	12.7
	B1	34	23.9
	B2	66	46.5
	C1	21	14.8
	Do not know	3	2.1
Reported previous AI-detector flagging	No	83	58.5
	Not sure	18	12.7
	Yes	41	28.9
Personal detector checking	Never	50	35.2
	Once	33	23.2
	More than once	59	41.5
Categorical humaniser intention	No	53	37.3
	Not sure	23	16.2
	Yes	66	46.5

Note. N = 142. Percentages may not total 100 because of rounding.

4.1. Levels and Most Commonly Reported Concerns

Tables 2 and 3 address RQ1 and RQ2. AI-detection anxiety was the most strongly endorsed construct and exceeded the scale midpoint. Humaniser intention was slightly above the midpoint, whereas writing self-censorship was approximately at it. All scales showed acceptable internal consistency.

All five anxiety items received majority agreement. The leading concerns were possible detector misclassification despite rule compliance and difficulty proving authorship. Choosing simpler vocabulary was the most endorsed self-censorship response, while the leading humaniser item concerned reducing the perceived risk that self-written text would be misclassified. Anxiety was therefore more prevalent than restrictive writing or prospective humaniser use.

Table 2: Item-Level Agreement and Descriptive Statistics

Item and statement	Valid *n*	*M*	*SD*	A/SA %
ADA1. I worry that English writing I produced myself could be flagged as AI-generated.	142	3.66	1.28	57.0
ADA2. I feel tense when submitting English writing that may be checked by an AI-writing detector.	142	3.64	1.27	59.9
ADA3. I worry that formal or polished English may make my writing appear AI-generated.	141	3.50	1.30	52.8
ADA4. I am concerned that I would have difficulty proving that the writing was my own if it were flagged.	142	3.66	1.27	61.3
ADA5. Even when I follow the assignment rules, the possibility of false AI detection makes me uneasy.	141	3.80	1.23	64.8
WSC1. I sometimes choose simpler vocabulary to reduce the chance that my writing will be flagged as AI-generated.	142	3.21	1.32	43.7
WSC2. I avoid sentence structures that seem too polished because they might appear AI-generated.	141	2.88	1.34	31.7
WSC3. I rewrite sentences that I produced myself mainly to make them look less AI-generated.	142	2.99	1.40	38.0
WSC4. I remove or replace expressions that seem natural to me because I fear they may trigger an AI-writing detector.	141	2.89	1.41	35.9
WSC5. I would accept less sophisticated English if I believed that doing so would reduce the risk of false flagging.	142	2.86	1.32	35.9
HI1. If my own writing received a high AI-detection score, I would consider using a humanizer tool before submission.	142	3.16	1.31	43.0
HI2. I would be willing to use a humanizer tool to reduce the chance that my self-written text would be falsely flagged.	142	3.30	1.32	47.2
HI3. If I believed that my institution relied heavily on detector scores, I would be likely to look for a humanizer tool.	141	2.94	1.36	36.6
HI4. I would consider using a humanizer tool even when the ideas and original draft were my own.	142	3.01	1.40	44.4
HI5. If I continued to fear false flagging, I would consider using a humanizer tool before submitting the text unchanged.	141	3.26	1.39	45.1

Note. Responses ranged from 1 (*strongly disagree*) to 5 (*strongly agree*). Valid *n* indicates the number of non-missing responses for each item. *M* = mean; *SD* = standard deviation; A/SA % = percentage of participants selecting *agree* or *strongly agree*; ADA = AI-detection anxiety; WSC = writing self-censorship; HI = humaniser intention. Percentages for items with one missing response were

calculated using the valid item sample. The spelling *humanizer* is retained only in the reproduced wording of the administered questionnaire items; *humaniser* is used elsewhere in the manuscript.

Table 3: Reliability and Construct-Level Descriptive Statistics

Construct	Items	Cronbach's α	*M*	*SD*
AI-detection anxiety	5	.835	3.65	0.98
Writing self-censorship	5	.766	2.97	0.98
Humaniser intention	5	.799	3.13	1.01

Note. $N = 142$. *M* = mean; *SD* = standard deviation. Construct scores ranged from 1 to 5; higher scores indicated stronger endorsement. Alpha coefficients were calculated from pairwise available covariances.

4.2. Experience Differences by Reported Previous AI-Detector Flagging

RQ3 compared the 41 learners reporting previous flagging with the 83 learners reporting no previous flagging. As shown in Table 4, learners reporting previous flagging reported significantly higher AI-detection anxiety. The effect was small to moderate. Writing self-censorship was also descriptively higher in this group, but the difference was not statistically significant and its effect size was small.

Table 4: Comparison by Reported Previous AI-Detector Flagging

Outcome	No reported flagging *M* (*SD*)	Reported flagging *M* (*SD*)	Mean difference [95% CI]	Welch's *t*	df	*p*	Cohen's *d* [95% CI]
AI-detection anxiety	3.54 (0.97)	3.92 (0.87)	0.38 [0.04, 0.73]	2.22	87.52	.029	0.41 [0.03, 0.79]
Writing self-censorship	2.88 (0.95)	3.04 (0.97)	0.16 [−0.21, 0.53]	0.87	78.61	.388	0.17 [−0.21, 0.54]

Note. Welch's independent-samples *t*-tests were used. *M* = mean; *SD* = standard deviation; CI = confidence interval; df = degrees of freedom. Mean differences and effect sizes are oriented as reported flagging minus no reported flagging. The 18 participants selecting not sure about previous flagging were excluded.

4.3. Reported Previous AI-Detector Flagging and Categorical Humaniser Intention

RQ4 examined whether reported previous AI-detector flagging was associated with categorical humaniser intention. Table 5 shows that affirmative intention was descriptively more common among learners reporting previous flagging. However, the overall association was not statistically significant, $\chi^2(2, N = 124) = 3.17$, $p = .205$, and the effect was small, Cramér's $V = .16$. The minimum expected cell count was 6.94, so the expected-frequency assumption was satisfied.

Table 5: Association Between Reported Previous AI-Detector Flagging and Categorical Humaniser Intention

Reported previous AI-detector flagging	No, n (%)	Not sure, n (%)	Yes, n (%)	Total n
No reported flagging	33 (39.8)	14 (16.9)	36 (43.4)	83
Reported flagging	10 (24.4)	7 (17.1)	24 (58.5)	41
Total	43 (34.7)	21 (16.9)	60 (48.4)	124

Note. Percentages in the two experience-group rows are row percentages. $\chi^2(2, N = 124) = 3.17$, $p = .205$, Cramér's $V = .16$. The 18 participants selecting not sure about previous AI-detector flagging were excluded. Humaniser intention denotes stated intention, not verified tool use.

5. Discussion and Limitations

5.1. Reported Levels of the Three Constructs

The RQ1 pattern indicates that apprehension about detector judgement was more prominent than either deliberate restriction of writing choices or willingness to use a humaniser. This ordering supports the conceptual separation of the three constructs.

Control-value theory helps interpret this pattern because anxiety is associated with important outcomes that are perceived as difficult to control (Pekrun, 2006, 2024). Writing affects grades and integrity judgements, while detector logic and appeal procedures may be unclear. Variation across tools and potential disadvantage for non-native writers make this uncertainty educationally consequential (Liang et al., 2023; Perkins et al., 2024; Weber-Wulff et al., 2023).

The midpoint-level self-censorship score further indicates that elevated anxiety did not uniformly correspond to restrictive writing. The three response patterns should therefore be interpreted as related but distinct dimensions of learners' experiences.

5.2. Most Commonly Reported Concerns

RQ2 identified the strongest concerns as possible detector misclassification and difficulty proving authorship. This pattern is consistent with evidence of variable detector performance and potential disadvantage for non-native English writing (Liang et al., 2023; Perkins et al., 2024; Weber-Wulff et al., 2023). Because detector outputs were not independently verified and the survey did not measure formal accusations, the finding represents perceived vulnerability rather than verified false-positive or accusation rates.

Choosing simpler vocabulary was the leading self-censorship response. The other items captured avoidance of polished sentence structures, rewriting of self-authored sentences, replacement of natural expressions, and willingness to accept less sophisticated English. In this Vietnamese EFL sample, these responses are relevant to vocabulary choice, sentence-structure choices, and the development of academic voice. If detector-oriented simplification became habitual, it could potentially reduce opportunities to practise more varied or sophisticated written English. However, the survey did not measure changes in linguistic proficiency, confidence in using English, sentence complexity as a textual outcome, or longer-term

academic language development. These therefore remain possible implications rather than demonstrated effects.

The most endorsed humaniser item framed tool use as protection for self-written text. Research showing that textual modification can reduce detector performance establishes the technical feasibility of detector management (Perkins et al., 2024; Weber-Wulff et al., 2023). The present finding instead identifies a possible unintended consequence: learners may contemplate modifying authentic writing to reduce perceived misclassification risk.

5.3. Differences by Reported Previous AI-Detector Flagging

RQ3 showed higher AI-detection anxiety among learners reporting previous flagging, with a small-to-moderate effect. This pattern is consistent with control-value theory, which conceptualises anxiety in relation to perceived control over valued outcomes (Pekrun, 2006, 2024).

Because the design was cross-sectional, directionality cannot be inferred. Learners who were already more concerned may have checked detectors more often, remembered ambiguous outputs more readily, or interpreted uncertain detector feedback as a flagging experience. The study also did not identify the detector, threshold, assessment context, appeal outcome, or whether any AI assistance occurred.

Writing self-censorship did not differ significantly between learners reporting previous flagging and those reporting no previous flagging, and the effect size was small. Thus, the statistically higher anxiety reported by learners with previous flagging experience was not accompanied by statistically detectable higher self-censorship. This pattern is consistent with treating anxiety and self-censorship as distinct constructs rather than assuming that greater evaluative concern necessarily co-occurs with restrictive changes in writing. However, the non-significant difference should not be interpreted as evidence that the groups were equivalent; forms of self-censorship not captured by the five-item measure may also have occurred.

5.4. Reported Previous AI-Detector Flagging and Humaniser Intention

RQ4 found no statistically significant association between reported previous AI-detector flagging and categorical humaniser intention. Although affirmative intention was descriptively more common among learners reporting previous flagging, the small effect did not provide sufficient evidence of an association.

Humaniser intention may reflect several untested considerations, including policy clarity, detector literacy, ethical judgement, familiarity with tools, and perceived reliability. The categorical measure should be interpreted as prospective willingness rather than observed humaniser use. Prospective willingness to consider detector-management strategies was present in both groups rather than being confined to learners reporting previous flagging.

5.5. Educational Implications

Taken together with prior evidence on detector limitations, the present findings support caution in treating AI-detector scores as stand-alone evidence of misconduct. Existing research on variable detector performance, vulnerability to textual modification, and potential inequity for non-native writers supports the use of informed human review, contextual evidence, and transparent opportunities to contest classifications (Liang et al., 2023; Perkins et al., 2024; Weber-Wulff et al., 2023). The present survey did not test these procedures or their effectiveness. Accordingly, clearer policies on permitted AI assistance, disclosure, detector use, and authorship review, together with proportionate process evidence such as drafts, revision histories, planning notes, or brief authorship conversations, should be understood as recommendations informed by the present findings and existing literature rather than as interventions evaluated in this study (Moorhouse et al., 2023).

For EFL instruction, the findings suggest the importance of addressing detector-related anxiety without encouraging learners to simplify legitimate language merely to appear less AI-generated. Teachers may help by explaining detector limitations, distinguishing ordinary revision from detector-oriented rewriting, and reinforcing that lexical range, sentence complexity, and academic voice should be developed for communicative and academic purposes rather than adjusted to satisfy an uncertain detector. Because confidence and longer-term language development were not measured, these pedagogical implications should be treated as cautious recommendations rather than observed outcomes of the present study.

5.6. Limitations

The single-site convenience sample limits generalisability, and the findings should not be assumed to represent Vietnamese EFL learners as a whole. All measures were self-reported, previous AI-detector flagging was not independently verified, and recall or social-desirability effects may have influenced sensitive responses. The cross-sectional design precludes temporal or causal inference. The three five-item measures were developed for this study and did not undergo formal expert review or pilot testing before administration. Their internal consistency was acceptable in the present sample, but factor structure, test–retest reliability, and broader construct validity were not examined. Excluding the 18 participants who were unsure about previous flagging reduced the comparative sample to 124. The study also did not examine detector brands, writing samples, policy clarity, or the circumstances surrounding reported flagging.

6. Conclusions and Future Work

This study found that AI-detection anxiety was more prominent than writing self-censorship or humaniser intention among 142 Vietnamese EFL learners. The most common concerns centred on possible detector misclassification despite rule compliance and difficulty proving authorship. Learners reporting previous flagging reported higher anxiety with a small-to-moderate effect, but they did not report significantly greater self-censorship, and reported flagging experience was not significantly associated with categorical humaniser intention.

The study contributes a learner-centred quantitative account that distinguishes emotional apprehension, restriction of legitimate writing, and prospective detector-management behaviour. Detector-related evaluation concerns can form part of learners' academic-writing experience even when misconduct is not established. In conjunction with prior evidence on detector limitations, the findings support cautious use of detector scores and suggest the importance of clear policies, proportionate authorship evidence, human review, and fair opportunities to contest classifications.

Future research should validate the scales in larger multisite samples and examine policy clarity, detector literacy, verified flagging, and actual revision behaviour. Longitudinal and mixed-method designs could clarify how concerns develop over time and whether they influence writing choices or tool use.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Quoc Khanh Nguyen: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft.

Nhat Quang Vo: Conceptualization, Formal analysis, Software, Validation, Writing – review and editing.

Both authors read and approved the final manuscript.

Funding

This research received no external funding.

Acknowledgments

The authors sincerely thank all participants who voluntarily completed the survey. Their time and responses made this study possible.

Ethical Statements

All participants provided informed consent before completing the anonymous, voluntary survey. No directly identifying information was collected, and findings were reported only in aggregate form. The absence of prospective ethics review and institutional permission is acknowledged as a procedural limitation.

Data and Code Availability

The de-identified data and R analysis code may be made available by the corresponding author upon reasonable request, subject to ethical and institutional considerations.

References

Büchi, M., Festic, N., & Latzer, M. (2022). The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. *Big Data & Society, 9*(1), 1–14. https://doi.org/10.1177/20539517211065368

Bui, T. T. U., & Tong, T. V. A. (2025). The impact of AI writing tools on academic integrity: Unveiling English-majored students' perceptions and practical solutions. *AsiaCALL Online Journal, 16*(1), 83–110. https://doi.org/10.54855/acoj.251615

Cong-Lem, N., Tran, T. N., & Nguyen, T. T. (2024). Academic integrity in the age of generative AI: Perceptions and responses of Vietnamese EFL teachers. *Teaching English with Technology, 24*(1), 28–47. https://doi.org/10.56297/FSYB3031/MXNB7567

Goddiksen, M. P., Johansen, M. W., Armond, A. C. V., Centa, M., Clavien, C., Gefenas, E., Kovács, N., Merit, M. T., Olsson, I. A. S., Poškutė, M., Santos, J. B., Santos, R., Strahovnik, V., Varga, O., Wall, P. J., Sandøe, P., & Lund, T. B. (2024). The dark side of text-matching software: Worries and counterproductive behaviour among European upper secondary school and bachelor students. *International Journal for Educational Integrity, 20*(1), Article 15. https://doi.org/10.1007/s40979-024-00162-7

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

Meniado, J. C., Huyen, D. T. T., Panyadilokpong, N., & Lertkomolwit, P. (2024). Using ChatGPT for second language writing: Experiences and perceptions of EFL learners in Thailand and Vietnam. *Computers and Education: Artificial Intelligence, 7*, Article 100313. https://doi.org/10.1016/j.caeai.2024.100313

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open, 5*, Article 100151. https://doi.org/10.1016/j.caeo.2023.100151

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R. (2024). Control-value theory: From achievement emotion to a general theory of human emotions. *Educational Psychology Review, 36*(3), Article 83. https://doi.org/10.1007/s10648-024-09909-7

Perkins, M., Roe, J., Vu, B. H., Postma, D., Hickerson, D., McGaughran, J., & Khuat, H. Q. (2024). Simple techniques to bypass GenAI text detectors: Implications for inclusive education. *International Journal of Educational Technology in Higher Education, 21*(1), Article 53. https://doi.org/10.1186/s41239-024-00487-w

Thi, X. H. N., Thien, H. V. H., Vuong, K. N., & Nguyen, T. T. (2025). Enhancing writing skills through AI-powered tools: Perceived benefits and challenges among Vietnamese EFL students. *Discover Education, 4*(1), Article 472. https://doi.org/10.1007/s44217-025-00905-9

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*(1), Article 26. https://doi.org/10.1007/s40979-023-00146-z