ARE AI DETECTORS FAIR TO EFL LEARNERS? A LIBRARY RESEARCH ON DETECTION BIAS AND ACADEMIC INTEGRITY

Jusak Patty [1] , Marles Yohannis Matatula [2]
[1] English Education Study Program
[2] English Education Postgraduate Study Program
Universitas Pattimura, Indonesia
Email: [1]jusak.patty@gmail.com, [2]matatulamarles@gmail.com

ABSTRACT

AI detection tools such as Turnitin AI Detection, GPTZero, and Copyleaks are now embedded in university assessment workflows to identify machine generated text, yet their performance on writing produced by non-native English speakers remains poorly understood. This critical library research synthesizes peer-reviewed literature published between 2022 and 2026 to examine three interrelated questions: how AI detectors perform on non-native English writing, what linguistic mechanisms produce detection bias against EFL learners, and what pedagogical and institutional implications follow from deploying these tools in linguistically diverse contexts. Literature was identified through Consensus AI using six targeted queries with filters restricting results to empirical, peer-reviewed studies. The review reveals that false positive rates for non-native writers exceed 61% in controlled studies, a disparity rooted in the overlap between the statistical profile of EFL writing and the features that detectors associate with machine generated text. Lexical predictability and syntactic uniformity, which are developmental characteristics of second language writing, produce the same low perplexity and uniform burstiness that detectors treat as markers of AI authorship. The uncritical adoption of these tools risks generating unjust accusations that compound existing inequities for EFL learners. Institutions should supplement automated detection with informed human review, redesign assessment toward process-oriented approaches, and develop context-sensitive policies that account for the linguistic profiles of second language writers.

Keywords: *AI detection tools, academic integrity, EFL writing, non-native English writers, false positives*

INTRODUCTION

Seven widely used GPT detectors, when tested on TOEFL essays written under supervised examination conditions, classified more than 61% of those essays as AI generated while identifying native English speaker essays with near-perfect accuracy (Liang et al., 2023). This disparity has been replicated across different detection platforms, linguistic populations, and testing conditions (Giray, 2024;

Hadra et al., 2026; Pratama, 2025), indicating that AI detection operates with a systematic performance gap tied to the language background of the writer rather than to any property of a single tool or dataset. The finding raises a question that the detection literature has only recently begun to address: whether the statistical logic embedded in current detection methodology is compatible with fair assessment of writing produced by speakers of English as a foreign language.

EFL writers compose under the cognitive and linguistic constraints that second language acquisition research has documented for decades, and these constraints produce text with statistical properties that overlap with the features detectors associate with machine authorship. The overlap is not coincidental but structural, rooted in how detection tools define the boundary between human and machine text and in whose writing that definition was built to recognize. The consequences of this design choice extend beyond technical accuracy into the domains of student welfare, institutional trust, and educational equity.

This study undertakes a critical literature review examining three dimensions of the problem: the documented performance of AI detection tools on non-native English writing, the linguistic and technical mechanisms through which detection bias against EFL writers originates, and the pedagogical and institutional implications of deploying these tools in linguistically diverse assessment contexts. The review draws on peer-reviewed literature published between 2022 and 2026, supplemented by foundational references from second language acquisition theory.

LITERATURE REVIEW

Perplexity, Burstiness, and the Logic of AI Text Detection

Commercial AI detection tools share a common inferential logic regardless of their proprietary architectures. Detection operates through two principal statistical measures. Perplexity quantifies how predictable a text's word choices are within a given linguistic context: a text composed of statistically expected word sequences yields low perplexity, while a text containing less expected selections yields high perplexity. Burstiness captures variation in sentence length and structural complexity across a document (Habibzadeh, 2023). Because language

models generate text by selecting the most probable next token at each step, their output tends to exhibit consistently low perplexity and uniform burstiness. Detection tools treat this statistical profile as a signature of machine authorship. A second category of detectors employs supervised machine learning classifiers trained on labeled corpora of human and AI text to identify distributional differences between the two categories (Hua & Yao, 2024), though most commercial platforms combine elements of both approaches. The critical assumption embedded across these methods is that human writing will deviate from the statistical regularity of machine output because human cognition introduces variability in word choice, sentence construction, and rhetorical strategy. This assumption establishes native English speaker writing as the implicit baseline for what human text looks like.

The reliability of this inferential logic depends on the representativeness of the baseline against which new texts are evaluated. Supervised classifiers learn the boundary between human and machine text from their training datasets, meaning that the composition of those datasets directly shapes what the classifier treats as typical human writing. If training corpora consist predominantly of texts produced by native English speakers, the classifier encodes the statistical profile of native speaker writing as the human norm and treats departures from that norm as evidence of machine generation. The training data of commercial detection tools remain proprietary and closed to independent audit, which means the extent to which linguistically diverse writing is represented cannot be verified externally (Liang et al., 2023). This opacity is consequential because it prevents institutions from evaluating whether a given detection tool has been validated on populations comparable to their own student body. Sadasivan et al. (2023) further demonstrated that the statistical markers on which detection relies are progressively eroded by improvements in language model sophistication, suggesting that the technical foundation of current detection methodology faces structural limitations that extend beyond any single population or tool.

Interlanguage Theory and the Complexity-Accuracy-Fluency Framework

Second language acquisition research offers a well-established theoretical account of the linguistic features that characterize EFL writing. Selinker (1972) introduced the concept of interlanguage to describe the systematic, rule-governed language system that learners construct during the process of acquiring a second language. Interlanguage is not a deficient approximation of the target language but a developing variety with its own internal consistency, shaped by first language transfer, learning strategies, and communicative demands. One of the empirically observable properties of interlanguage writing is its constrained lexical and syntactic profile. EFL writers tend to employ a narrower vocabulary range, favor high-frequency items, and produce lower scores on established lexical diversity measures compared to more proficient or native speakers (Crossley, 2020; Kyle et al., 2021). Furthermore, while advanced learners eventually develop sophisticated phraseological patterns, lower-level interlanguage writing often lacks this phraseological sophistication and associative strength (Paquot, 2019).

The complexity-accuracy-fluency (CAF) framework provides a theoretical explanation for these patterns. Originally proposed by Skehan (1998) and subsequently refined by Housen & Kuiken (2009), CAF theory posits that second language writers operate under limited attentional capacity and must distribute cognitive resources across competing dimensions of performance. When composing in their L2, writers frequently prioritize accuracy or communicative clarity at the expense of advanced syntactic elaboration. Rather than utilizing native-like phrasal compression, this trade-off often results in an overreliance on rigid clausal subordination and less structural flexibility (Kuiken, 2023; C. Zhang & Kang, 2022). Skehan (2009) extended this framework to include lexical complexity as a fourth dimension, further specifying how the trade-off operates across multiple levels of linguistic production. The resulting text exhibits a statistical profile with reduced predictability and uniform burstiness, not because the writer lacks linguistic knowledge, but because the cognitive demands of second language production constrain the range of productive resources deployed in real-time composition. These theoretical predictions regarding syntactic and lexical

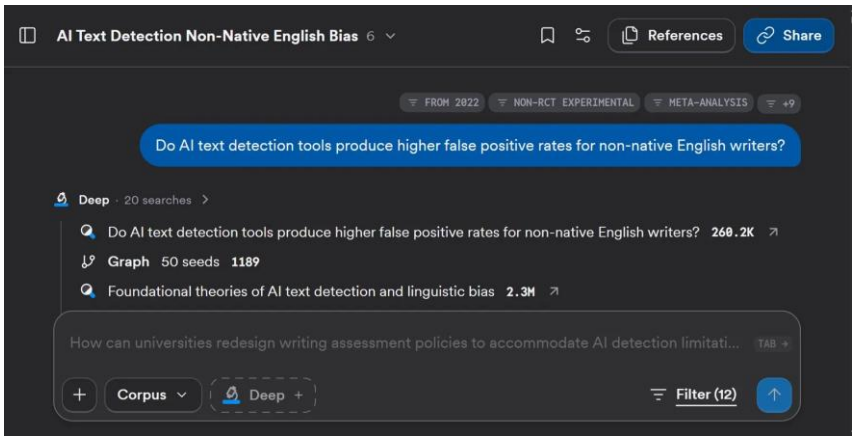
profiles have been consistently supported by corpus-based studies of EFL academic writing (Y. Zhang & Ouyang, 2023).

RESEARCH METHOD

This study adopts a library research design using the critical review methodology described by Grant & Booth (2009). A critical review evaluates and interprets available evidence to develop a conceptual contribution rather than to catalog every relevant study. This approach suits the present topic because the research questions require synthesis across computational linguistics, second language acquisition, and educational assessment, and because the emerging nature of the literature demands interpretive analysis rather than statistical aggregation (Snyder, 2019).

Literature was identified through Consensus AI (consensus.app), an academic search engine indexing over 200 million peer-reviewed papers. The search used the platform's deep analysis mode with 12 filters: publication year from 2022 to 2026, preprints excluded, empirical study designs (systematic review, meta-analysis, non-randomized controlled trial, and observational study), human studies only, and field of study restricted to Education, Linguistics, Computer Science, and Psychology. Six targeted queries were formulated to address each research question from complementary angles: (1) "Do AI text detection tools produce higher false positive rates for non-native English writers?" (2) "How accurate are AI detection tools such as Turnitin and GPTZero on EFL student writing?" (3) "How do lexical diversity and syntactic complexity of EFL writing affect AI detection accuracy?" (4) "Why do AI text detectors misclassify non-native English writing as AI generated?" (5) "What are the consequences of AI detection false positives for students in higher education?" (6) "How should universities address AI detection bias against non-native English speakers?" Figure 1 illustrates the search interface and filter configuration used in the platform.

Figure 1. Literature Search Configuration Using Consensus AI Deep Analysis Mode



Sources were included if they were peer-reviewed, published in English between 2022 and 2026, and addressed the performance, bias, or implications of AI detection tools in relation to non-native English writing. Sources were excluded if they were preprints, focused solely on algorithm development without educational implications, or addressed traditional plagiarism detection without a generative AI component. Additional references were identified through backward snowballing, and foundational works in second language acquisition theory were included where the analysis required theoretical grounding. After deduplication and screening, 38 sources formed the final corpus. Analysis followed a thematic synthesis organized around the three research questions, prioritizing integration across studies over sequential reporting.

FINDINGS AND DISCUSSION

Detection Performance on Non-Native English Writing

Controlled testing reveals a performance gap large enough to call into question whether current AI detectors can be used for consequential decisions in contexts that include non-native English writers. False positive rates for non-native essays exceed 61% under supervised examination conditions where AI assistance is impossible, while native speaker essays in the same studies are identified correctly with near-perfect accuracy (Liang et al., 2023). This disparity has been confirmed across different detection platforms, testing conditions, and linguistic

populations (Gotoman et al., 2025; Lege, 2025; Pratama, 2025), establishing that the problem is not confined to a single tool or dataset but reflects a pattern embedded in how detection technology operates as a category. The disparity also appears to follow a proficiency gradient: essays flagged unanimously by all detectors in the Liang et al.'s (2023) study exhibited significantly lower perplexity than those flagged by fewer detectors, suggesting that writers whose productive linguistic repertoire is most constrained face the highest risk of misclassification. In university contexts where student English proficiency spans a wide range, this gradient means that the students most likely to be falsely accused are those who are already most educationally vulnerable.

Cross-platform inconsistency deepens the concern. Independent evaluations consistently report that the same text submitted to different detectors yields contradictory verdicts, that many commercially available tools fail to achieve 80% accuracy despite exceptions among a few premium platforms, and that vendor accuracy claims routinely exceed observed performance under controlled testing (Elkhatat et al., 2023; Walters, 2023; Weber-Wulff et al., 2023). For EFL assessment contexts specifically, one recent evaluation concluded that current tools are unsuitable as authoritative indicators of authorship (Hadra et al., 2026). The practical consequence of this inconsistency is that outcomes for individual students may depend less on what they wrote than on which detection platform their institution happens to have licensed. When a tool whose accuracy falls below the threshold required for reliable decision-making is nonetheless used to initiate academic misconduct proceedings, the evidential basis for those proceedings is fundamentally compromised.

A partial counterpoint deserves acknowledgment. Detection models trained on balanced datasets that include both native and non-native writing samples have demonstrated substantially reduced bias without major loss in overall accuracy (Ibrahim et al., 2025; Jiang et al., 2024). These results indicate that the bias observed in current commercial tools is not an inherent limitation of all detection technology but a consequence of design choices, particularly the composition of training data and the weighting of statistical features. The practical reach of this

finding, however, remains limited. The tools most widely deployed in university assessment workflows are commercial platforms whose training data remain proprietary, and the independent testing record shows that these platforms have not yet achieved the performance parity that experimental models demonstrate under controlled conditions (Giray, 2024). Until commercial detectors incorporate the design improvements that experimental research has shown to be effective, the bias against non-native writers persists as a feature of the tools that institutions actually use rather than the tools that researchers have shown to be possible.

Linguistic Mechanisms of Detection Bias

The performance gap documented above raises a question that the detection literature alone cannot answer: why does non-native writing trigger false positives at such elevated rates? The answer lies in the complex statistical relationship between EFL writing and AI-generated text. Direct comparisons of L2 student essays and ChatGPT-generated essays reveal that while AI vastly outperforms developing writers on syntactic complexity measures like dependent clauses per T-unit (Fredrick & Craven, 2025), there is partial convergence on lexical diversity measures in specific academic tasks (Wu, 2025). The overlap, however, is not total: at the level of collocation and lexical patterning, human EFL writing and AI output differ in ways that reflect the distinct processes generating each text type (M. Zhang & Crosthwaite, 2025). But these fine-grained differences are precisely the differences that current detection tools do not measure. The tools operate on aggregate statistical features where the two text types can produce comparable metric signatures, not on the qualitative dimensions where they diverge.

Multidimensional and large-scale analyses reinforce this finding by highlighting the structural asymmetry in detection methodology. AI-generated text tends to outperform student writing on lexical and syntactic sophistication measures while lacking rhetorical hedging, personal voice, and contextual adaptability (Herbold et al., 2023). Multidimensional analysis reveals significant disparities between language model output and actual human registers, indicating that AI fails to authentically emulate the spontaneous nuances of human texts (Berber Sardinha,

2024). Furthermore, examination of verb collocations shows that EFL writers actually produce more diverse and context-specific collocations than AI, whereas AI output is characterized by higher lexical repetition but near-perfect grammatical accuracy (Du et al., 2025). Linguistic fingerprinting can distinguish ChatGPT from EFL prose at these fine-grained levels (Mizumoto et al., 2024), but the features required for this distinction are not those that commercially deployed detectors currently assess.

A useful conceptual distinction for understanding this asymmetry comes from corpus research comparing high-scoring human IELTS essays with AI generated equivalents (Wu, 2025). Human essays that scored well tended to exhibit what can be characterized as deep logical complexity: nested dependent structures deployed for precise argumentation, with subordination serving a specific rhetorical function within the argument. AI output, by contrast, tended toward breadth complexity: expanding surface structures and multiplying coordinate clauses without matching the depth of logical organization. Both types of text, however, produced comparable scores on the aggregate measures that detectors use. This distinction captures the core of the problem. Detection tools cannot differentiate between complexity that arises from genuine intellectual engagement with a topic and regularity that arises from probabilistic optimization, because both produce similar statistical signatures at the level of measurement that detectors operate on. For EFL writers, this means that their developmental linguistic constraints place them in a detection blind spot where their authentic writing is statistically indistinguishable from machine output on the specific dimensions that determine classification.

Pedagogical and Institutional Implications

False accusations carry consequences that persist beyond the resolution of any individual case. Students subjected to AI-based misconduct investigations report experiencing hostility, anxiety, fear, and defeat, alongside a lasting erosion of trust in both the accusing instructor and the institution (Crockett & Howe, 2024; Gorichanaz, 2023). Some students even expressed profound disillusionment with the higher education system itself as a result of these experiences (Gorichanaz,

2023). These psychological effects are not contingent on the outcome of the investigation; the experience of being accused is itself damaging, regardless of whether the accusation is ultimately sustained or withdrawn. For non-native English speakers, the harm is compounded by a structural inequality that operates prior to and independent of any individual detection event: they must already navigate higher barriers in English-medium academic environments, and the added risk of false accusation introduces a layer of disadvantage that no amount of individual resilience can compensate for (Wee & Reimer, 2023). When an institutional mechanism intended to protect integrity generates systematic harm for a particular demographic group, the mechanism itself becomes a source of inequity rather than a safeguard against it.

The literature points toward a set of institutional responses that together represent a more defensible approach than detection-centered enforcement. Automated flagging should initiate dialogue rather than formal proceedings, and the dialogue should be conducted by an educator with an understanding of the student's linguistic profile and writing trajectory (Deep et al., 2025; Giray et al., 2025). Assessment design should shift toward process-oriented approaches that evaluate writing development over time rather than evaluating finished products for signs of AI involvement (Kim & Danilina, 2025). Institutional resources currently directed toward detection and enforcement may be more productively invested in developing AI literacy among both educators and students, on the basis that teaching responsible use is more sustainable than attempting to detect and punish misuse (Holbeck, 2025). None of these recommendations requires abandoning concern for academic integrity; they require redirecting that concern toward approaches that do not carry the equity costs that automated detection imposes on linguistically diverse populations.

These recommendations acquire particular specificity in EFL contexts where the evidence from the region provides relevant data. Indonesian doctoral students who participated in a phenomenological study on AI use in academic writing described strategic engagement with AI tools accompanied by genuine ethical concern about the boundaries of acceptable assistance (Pratiwi et al., 2025).

Similarly, adult Indonesian EFL learners preparing for postgraduate studies showed affordance-oriented enthusiasm, noting that AI provided a less anxiety-provoking learning environment while they actively navigated ethical boundaries to avoid plagiarism (Arifin et al., 2025). What these studies reveal is a student population that is actively negotiating ethical questions about AI, not a population that requires technological surveillance. The disconnect between what detection tools assume about students and what research documents about their actual engagement with AI tools represents perhaps the most consequential gap in the current institutional response. Closing that gap requires policy frameworks that are developed with attention to the specific contexts in which they will be applied, rather than imported wholesale from English-dominant settings where the linguistic dynamics are fundamentally different (Farrelly & Baker, 2023).

CONCLUSION

This review finds that AI detection tools produce false positive rates exceeding 61% for non-native English writing under controlled conditions, a disparity rooted in the structural overlap between the statistical profile of EFL writing and the features that detectors associate with machine generated text. The deployment of these tools in EFL contexts generates documented psychological harm and compounds existing educational inequities for non-native speakers. Institutions serving linguistically diverse populations should treat detection flags as starting points for informed human review rather than as evidence of misconduct, and should invest in process-oriented assessment and AI literacy as more equitable alternatives to detection-centered enforcement. This study is limited by its reliance on secondary literature and by the rapid evolution of detection technology, which may render specific accuracy figures provisional. No large-scale empirical study has yet evaluated detection bias on Indonesian EFL writing corpora, and future research addressing this gap would contribute meaningfully to understanding the contextual specificity of the problem.

REFERENCES

Arifin, M. A., Rahman, A. A., Balla, A., Susanto, A. K., & Pratiwi, A. C. (2025). ChatGPT Affordances and Indonesian EFL Students' Perceptions in L2 Writing: A Collaborative Reflexive Thematic Analysis. *Changing English*, *32*(2), 195–211. https://doi.org/10.1080/1358684X.2024.2418132

Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, *4*(1), 100083. https://doi.org/10.1016/j.acorp.2023.100083

Crockett, R., & Howe, R. (2024). The Inherent Uncertainties of AI-Text Detection and the Implications for Education Institutions: An Overview. In S. Mahmud (Ed.), *Advances in Educational Marketing, Administration, and Leadership* (pp. 175–198). IGI Global. https://doi.org/10.4018/979-8-3693-0240-8.ch010

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(vol. 11 issue 3), 415–443. https://doi.org/10.17239/jowr-2020.11.03.01

Deep, P. D., Edgington, W. D., Ghosh, N., & Rahaman, Md. S. (2025). Evaluating the Effectiveness and Ethical Implications of AI Detection Tools in Higher Education. *Information*, *16*(10), 905. https://doi.org/10.3390/info16100905

Du, M., Lu, M., Dai, Y., & Wang, F. (2025). A Corpus-Based Analysis of Verb Collocations in Human and AI-Generated IELTS Writing. *Journal of Educational Technology and Innovation*, *7*(2). https://doi.org/10.61414/mzmmrq74

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Farrelly, T., & Baker, N. (2023). Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice. *Education Sciences*, *13*(11), 1109. https://doi.org/10.3390/educsci13111109

Fredrick, D. R., & Craven, L. (2025). Lexical diversity, syntactic complexity, and readability: A corpus-based analysis of ChatGPT and L2 student essays. *Frontiers in Education*, *10*, 1616935. https://doi.org/10.3389/feduc.2025.1616935

Giray, L. (2024). The Problem with False Positives: AI Detection Unfairly Accuses Scholars of AI Plagiarism. *The Serials Librarian*, *85*(5–6), 181–189. https://doi.org/10.1080/0361526X.2024.2433256

Giray, L., Sevnarayan, K., & Ranjbaran Madiseh, F. (2025). Beyond Policing: AI Writing Detection Tools, Trust, Academic Integrity, and Their Implications for College Writing. *Internet Reference Services Quarterly*, *29*(1), 83–116. https://doi.org/10.1080/10875301.2024.2437174

Gorichanaz, T. (2023). Accused: How students respond to allegations of using ChatGPT on assessments. *Learning: Research and Practice*, *9*(2), 183–196. https://doi.org/10.1080/23735082.2023.2254787

Gotoman, J. E. J., Luna, H. L. T., Sangria, J. C. S., Santiago Jr., C. S., & Barbuco, D. D. (2025). Accuracy and Reliability of AI-Generated Text Detection Tools: A Literature Review. *American Journal of IR 4.0 and Beyond*, *4*(1), 1–9. https://doi.org/10.54536/ajirb.v4i1.3795

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Habibzadeh, F. (2023). GPTZero Performance in Identifying Artificial Intelligence-Generated Medical Texts: A Preliminary Study. *Journal of Korean Medical Science*, *38*(38), e319. https://doi.org/10.3346/jkms.2023.38.e319

Hadra, M., Cambridge, K., & Mesbah, M. (2026). Evaluating the accuracy and reliability of AI content detectors in academic contexts. *International Journal for Educational Integrity*, *22*(1), 4. https://doi.org/10.1007/s40979-026-00213-1

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*(1), 18617. https://doi.org/10.1038/s41598-023-45644-9

Holbeck, R. (2025). Beyond Detection: Why faculty should focus on AI literacy, not AI policing. *eLearn*, *2025*(5), 3735548.3729174. https://doi.org/10.1145/3735548.3729174

Housen, A., & Kuiken, F. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Hua, H., & Yao, C.-J. (2024). Investigating generative AI models and detection techniques: Impacts of tokenization and dataset size on identification of AI-generated text. *Frontiers in Artificial Intelligence*, *7*, 1469197. https://doi.org/10.3389/frai.2024.1469197

Ibrahim, K. H. S., Al Otaibi, D., & Sibai, F. N. (2025). The Robustness of AI-Classifiers in the Face of AI-Assisted Plagiarism: The Case of Turnitin AI Content Detector. *International Journal of Computer-Assisted Language Learning and Teaching*, *15*(1), 1–27. https://doi.org/10.4018/IJCALLT.375428

Jiang, Y., Hao, J., Fauss, M., & Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers & Education*, *217*, 105070. https://doi.org/10.1016/j.compedu.2024.105070

Kim, J., & Danilina, E. (2025). Towards inclusive and equitable assessment practices in the age of GenAI: Revisiting academic literacies for multilingual students in academic writing. *Innovations in Education and Teaching International*, *62*(5), 1593–1597. https://doi.org/10.1080/14703297.2025.2456223

Kuiken, F. (2023). Linguistic complexity in second language acquisition. *Linguistics Vanguard*, *9*(s1), 83–93. https://doi.org/10.1515/lingvan-2021-0112

Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, *43*(4), 781–812. https://doi.org/10.1017/S0272263120000546

Lege, R. P. (2025). Auditing the Fairness of AI-Detection Tools: A Comparative Study of ESL, Published, and AI-Generated Texts and Their Misclassification Risks. *International Journal of Teaching, Learning and Education*, *4*(5), 30–45. https://doi.org/10.22161/ijtle.4.5.5

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

Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics*, *4*(3), 100106. https://doi.org/10.1016/j.acorp.2024.100106

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Pratama, A. R. (2025). The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science*, *11*, e2953. https://doi.org/10.7717/peerj-cs.2953

Pratiwi, H., Suherman, Hasruddin, & Ridha, M. (2025). Between Shortcut and Ethics: Navigating the Use of Artificial Intelligence in Academic Writing Among Indonesian Doctoral Students. *European Journal of Education*, *60*(2), e70083. https://doi.org/10.1111/ejed.70083

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023). *Can AI-Generated Text be Reliably Detected?* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2303.11156

Selinker, L. (1972). Interlanguage. *IRAL - International Review of Applied Linguistics in Language Teaching*, *10*(1–4). https://doi.org/10.1515/iral.1972.10.1-4.209

Skehan, P. (1998). *A cognitive approach to language learning* (9. [Dr.]). Oxford Univ. Press.

Skehan, P. (2009). Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Walters, W. H. (2023). The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Information Science*, *7*(1), 20220158. https://doi.org/10.1515/opis-2022-0158

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), 26. https://doi.org/10.1007/s40979-023-00146-z

Wee, H. B., & Reimer, J. D. (2023). Non-English academics face inequality via AI-generated essays and countermeasure tools. *BioScience*, *73*(7), 476–478. https://doi.org/10.1093/biosci/biad034

Wu, J. (2025). Comparing Linguistic Features between Human-written High-Scoring IELTS Essays and AI-Generated ones. *Journal of Humanities and Social Sciences Studies*, *7*(8), 68–77. https://doi.org/10.32996/jhsss.2025.7.8.8

Zhang, C., & Kang, S. (2022). A comparative study on lexical and syntactic features of ESL versus EFL learners' writing. *Frontiers in Psychology*, *13*, 1002090. https://doi.org/10.3389/fpsyg.2022.1002090

Zhang, M., & Crosthwaite, P. (2025). More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2024-0196

Zhang, Y., & Ouyang, J. (2023). Linguistic complexity as the predictor of EFL independent and integrated writing quality. *Assessing Writing*, *56*, 100727. https://doi.org/10.1016/j.asw.2023.100727