



Using Automated Writing Evaluation to Improve EFL Students’ IELTS Writing: Effects Across Proficiency Levels and Learning Behaviours

Napat Jitpaisarnwattana^{a,*}, Nick Saville^b

^a napat.jit@gmail.com, Silpakorn University, Thailand
^b nick.saville@cambridge.org, Cambridge University Press and Assessment
* Corresponding author, napat.jit@gmail.com

APA Citation: Jitpaisarnwattana, N., & Saville, N. (2026). Using automated writing evaluation to improve EFL students’ IELTS writing: Effects across proficiency levels and learning behaviours. <i>LEARN Journal: Language Education and Acquisition Research Network</i> , 19(2), 265-296. https://doi.org/10.70730/XYWK3308	
Received 27/05/2025	ABSTRACT This study examines the impact of Write & Improve, an Automated Writing Evaluation (AWE) tool, on Thai university students' performance in IELTS writing. Additionally, it explores learning behaviors that may contribute to writing development. Employing a quantitative approach, the study gathered data from 53 students enrolled in an English Preparation for Standardized Tests course at a Thai public university. The effectiveness of the AWE tool was assessed through a comparative analysis of pre- and post-test scores, while learning behaviors were examined using program-generated learning logs. The findings indicate that the Write & Improve intervention led to significant improvements in students’ post-test writing performance across both writing tasks and all assessment criteria, with the most notable gains observed in grammatical range and accuracy. The most substantial progress was evident among intermediate-level students (IELTS Band 5.5 and below). Furthermore, learning analytics suggests that students who engaged in a greater number of writing tasks demonstrated stronger post-test performance. The study concludes that while AWE tools can
Received in revised form 18/03/2026	
Accepted 31/03/2026	

	facilitate improvements in specific writing components, their integration should be guided by well-defined pedagogical objectives, with teacher involvement remaining essential in the feedback process. Keywords: Automated Writing Evaluation, IELTS writing, Write and Improve, automated feedback, learning analytics
--	--

Introduction

Automated writing evaluation (AWE), defined as the systematic assessment and scoring of learners' written output through computational algorithms (Shermis & Burstein, 2003), has attracted significant attention within the language assessment community. This interest extends across various stakeholders, including test developers, institutional administrators, researchers, and educators, highlighting its growing role in evaluating writing proficiency. Various stakeholders have leveraged AWE for diverse purposes, ranging from scoring high-stakes standardised assessments such as the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examination (GRE) to facilitating grading and feedback provision for low-stakes writing tasks in classroom settings (Zhang, 2021). Initially designed to generate summative scores for written texts, AWE systems have increasingly been repurposed to offer formative feedback on student writing. This shift allows learners to engage in iterative revision and rewriting processes, fostering writing development for pedagogical purposes (Cotos, 2014; Thuy, 2022). The transition from summative to formative assessment has been partly driven by evolving perspectives within the language assessment community, emphasizing the need to integrate learning as a fundamental component of assessment. This paradigm shift aligns with the notion that language assessment should be conducted 'for learning,' rather than merely measuring outcomes. Consequently, this reconceptualization has led to the emergence of Learning-Oriented Assessment (LOA), which is grounded in the principle of fostering learning by utilizing various forms of assessment to support and enhance the learning process (Jitpaisarnwattana & Saville, 2025).

Advancements in AWE technology, particularly through the integration of artificial intelligence (AI), have significantly influenced writing instruction across multiple dimensions (Wang et al., 2013). One key advantage is the immediacy of feedback; unlike teacher- or peer-provided feedback, AWE systems can deliver near-instantaneous feedback to learners regardless of time or location (Wang et al., 2013; Jitpaisarnwattana & Saville, 2025). Another notable benefit is the expansion of writing opportunities, as students can engage in writing tasks as frequently as desired, accompanied by

detailed feedback and error explanations. Lastly, the feedback and scores generated by AWE are often perceived as more ‘objective’ compared to those provided by human raters, particularly in assessing writing mechanics (Rezaei & Lovorn, 2010). Although AWE offers several advantages, its effectiveness remains a subject of concern, particularly regarding the accuracy of its feedback and its capacity to deliver substantive and pedagogically meaningful explanations beyond merely identifying grammatical errors (Fan, 2023; Shi et al., 2022).

Empirical research on AWE can be categorized into four primary strands. The first focuses on the reliability of AWE-generated scores in comparison to human raters (Ranalli et al., 2017; Shermis, 2014). The second examines AWE through the lens of validation frameworks (Chappelle et al., 2015). The third explores students’ perceptions and acceptance of AWE (Li et al., 2019; Wang, 2015; Zhai & Ma, 2022). The final research strand investigates the impact of AWE implementation on students’ writing development, revealing both positive and negative effects on writing performance, particularly in terms of accuracy, autonomy, and interaction (Chen & Cheng, 2008; Dikli, 2006; Wang et al., 2013; Ware & Warschauer, 2006). While existing studies predominantly examine the impact of AWE on overall writing development, recent advancements in AWE technology have enabled the assessment of specific writing genres, such as those required in standardised tests like IELTS and TOEFL. However, to the best of our knowledge, there remains a lack of research investigating the effects of AWE use on students’ writing development in the context of standardised test preparation

The current study, therefore, aims to investigate the effects of Write and Improve, an AWE program, on students’ IELTS writing based on the four IELTS criteria: task response (and task achievement), lexical resource, coherence and cohesion and grammatical range and accuracy and to investigate learning behaviours within the AWE program that may contribute to writing improvement. The study is guided by two primary research questions and one sub-research question:

- 1) What are the effects of Write and Improve on students’ IELTS writing based on the four IELTS criteria: task response (and task achievement), lexical resource, coherence and cohesion and grammatical range and accuracy?

- 1.1) Which group of students (intermediate upper-intermediate/ advanced) benefited the most from Write and Improve?

- 2) To what extent did learning behaviours displayed in Write and Improve contribute to improvements in writing performance in a post-test?

Literature Review

Automated Writing Evaluation

Automated Writing Evaluation (AWE) is recognized as a form of intelligent instructional technology, as well as an interdisciplinary domain of research and development within educational technology (Cotos, 2018). It integrates insights from multiple disciplines, including linguistics, second language acquisition (SLA), computer science, psychometrics, and educational measurement. AWE originated in the 1960s with the development of Project Essay Grade (PEG), which was introduced to address the increasing demands on teachers for grading student essays and to support large-scale testing programs. However, significant advancements in computer technology in the 1990s facilitated more sophisticated computational analyses of naturally occurring language and enabled the electronic processing of constructed written responses, thereby enhancing the capabilities of AWE systems (Cotos, 2018). These developments collectively led to the emergence of various AWE programs, including e-rater, Intelligent Essay Assessor (IEA), Intellimetric, and AutoScore, among others. However, during this period, AWE applications were primarily designed to score essays by replicating human raters' evaluative processes. Over time, these systems were integrated into large-scale standardized assessments, such as e-rater for the TOEFL, Intellimetric for the GMAT, and the Intelligent Essay Assessor for the Pearson Test of English (PTE).

Later generations of Automated Essay Scoring (AES) systems expanded their functionality beyond merely generating summative scores, shifting towards the provision of formative writing feedback (Li et al., 2015; Ranalli et al., 2017). Criterion®, WriteToLearn and MyAccess! emerged as enhanced iterations of e-rater, Intelligent Essay Assessor and Intellimetric, respectively. This change in focus and capabilities led to several terminological shifts, transitioning from Automated Essay Scoring (AES) to Automated Essay Evaluation (AEE), and ultimately to the more encompassing term, Automated Writing Evaluation (AWE). This generation of AWE systems prioritized comprehensive writing assessment, extending beyond essay evaluation to other writing genres. These systems were designed to analyze writing drafts and provide feedback on multiple dimensions of writing, including grammatical accuracy, syntactic complexity, topic relevance, lexical resources, coherence, and discourse management, among other aspects (Ranalli & Hegelheimer, 2022).

Different Applications of AWE

AWE systems serve multiple functions, with two primary applications: (1) generating summative writing scores for standardized assessments and (2) providing formative feedback to support writing development in classroom settings. The early development of AWE primarily focused on essay scoring. PEG was designed to assess student essays previously graded by English teachers, while e-rater, developed by ETS, was initially used for the writing component of the GMAT before being adopted for the TOEFL. Developed by Vantage Learning, IntelliMetric employs NLP and statistical methods to assess writing sections in standardized tests (Zhang, 2021). Since 2006, it has been used for scoring the Analytical Writing Assessment in the GMAT. In addition, the Intelligent Essay Assessor, an AWE system, has been used for standardized assessment, including as a scoring mechanism for the Pearson Test of English (PTE). Unlike conventional AWE systems that focus on linguistic accuracy and structure, it prioritizes the conceptual and ideological depth of written responses (Tang & Wu, 2011). These automated scoring systems have been integrated into large-scale standardized assessments, including TOEFL, GMAT, and PTE. However, these systems do not function independently; rather, they are employed in conjunction with human raters to ensure a more comprehensive and reliable evaluation of writing proficiency.

Recent advancements in AWE have shifted toward learning-oriented applications, with more programs designed as instructional tools (Warschauer & Grimes, 2008). Systems originally developed for scoring standardized test essays have been adapted for classroom use, including ETS's Criterion®, Pearson's WriteToLearn, and Vantage Learning's MyAccess! These feedback-driven AWE systems represent a new generation focused more on learning than scoring. More recent AWE software, such as Write & Improve (W&I), adopts an even more learner-centered approach, incorporating a dashboard for tracking progress. Originally developed from the Cambridge Learner Corpus (CLC), it leverages written data from L2 English learners who participated in Cambridge English exams (Yanakudakis et al., 2018). Scores and feedback differ across programs, with some offering general writing feedback, while others provide targeted feedback for specific writing types. For instance, Criterion® provides detailed feedback on discourse structure, including thesis statements, supporting arguments, and conclusions, along with a numerical score (Warschauer & Grimes, 2008). Write & Improve detects errors in sentence structure, grammar, punctuation, and word choice, assigning scores aligned with CEFR and IELTS descriptors. It also provides hints and suggestions for correcting the errors or improving the choice of words.

Benefits and Challenges of AWE

While there is no consensus among researchers and practitioners on the effectiveness of AWE for writing development (Hockly, 2019), both theoretical discussions and empirical studies have highlighted its potential benefits. The key advantage of AWE is its time efficiency (Cotos, 2011). It provides instant scoring and feedback, particularly for surface-level errors like grammar, spelling, and punctuation. This allows teachers to focus more on higher-order writing aspects, such as content and argumentation (Woodworth & Barkaoui, 2020). Moreover, AWE enhances objectivity and consistency across classes and curricula (Thuy et al., 2022). Unlike human raters, who may be influenced by individual student differences, AWE provides standardized scoring and feedback, ensuring greater reliability (Parra & Calero, 2019). Another key advantage of AWE systems is their ability to enhance learners' awareness of language accuracy and error detection. The diagnostic feedback they provide helps learners recognize and correct mistakes independently based on the written corrective feedback (WCF) received (Chou & Chung, 2013; Jitpaisarnwattana, 2022, 2024; Ranalli et al., 2017). This process promotes greater attention to grammatical rules in WCF, supporting their independent application in future writing tasks (Zhang, 2017).

While AWE has gained traction in writing instruction and assessment, several inherent and practical challenges remain. The primary concern is the nature of AWE feedback, which can be vague and difficult to interpret. Often, it lacks concrete suggestions, particularly regarding content and rhetorical aspects of writing (Cheng, 2006; Thuy et al., 2022; Wang, 2015). Also, the accuracy of AWE feedback remains a concern, with errors sometimes overlooked or misclassified. Lavolette et al. (2015) found that only 75% of identified errors were accurate, while 14% were misclassified. This can lead to incorrect error notifications, potentially causing inaccurate learning and reluctance to use the feedback (Burstein, 2003; Woodworth & Barkaoui, 2020). Moreover, despite being instant and comprehensive, AWE feedback is often unfocused and non-adaptive. Most systems treat all errors equally, regardless of their impact on communication. Additionally, feedback is standardized, failing to consider learners' proficiency, goals, or readiness to engage with it (Woodworth & Barkaoui, 2020). Another challenge is the absence of social and communicative aspects in AWE. These programs can create unnatural writing environments that diminish human interaction (Herrington & Moran, 2001). Unlike real-world contexts, automated systems lack authentic interaction, meaningful communication, and rich negotiation of meaning (Ferris & Hedgcock, 2005; Ware & Warschauer, 2006).

Empirical Studies on AWE and Writing Development

Research on AWE has been explored from multiple perspectives. Early studies focused on comparing computer-generated and human scores, while more recent research examines its role in language learning. Previous research on the impact of AWE on student writing has produced mixed but predominantly positive results. El Ebyary and Windeatt (2010) investigated Criterion's impact on EFL trainee teachers in Egypt, finding that automated feedback enhanced writing accuracy. Participants also showed improved proficiency, reflected in higher AWE-generated scores in later submissions. In addition, Wang et al. (2013) investigated the effectiveness of CorrectEnglish, an AWE program, in enhancing the writing quality of university students in Taiwan. Their findings indicated a significant improvement in writing accuracy, with post-test results showing a reduction in errors related to run-on sentences, fragments, articles, and punctuation. Moreover, Liao (2016) investigated Criterion's impact on grammatical accuracy in revisions and new compositions, finding that AWE feedback effectively reduced run-on sentences, fragments, subject-verb agreement errors, and ill-formed verbs. The corrective feedback provided by Criterion also contributed to increased revisions and improved linguistic accuracy in both initial and final drafts (Li et al., 2015).

A few studies have explored the application of Write & Improve in language learning. Karpova (2020) integrated the tool into an English course at a university in Ukraine and found that students perceived it as beneficial for improving their writing skills. However, the findings were primarily based on self-reported perceptions, as no direct assessments of students' writing performance were conducted. On top of that, Liu & Yu (2022) employed eye-tracking technology, complemented by stimulated recall protocols and reflective journals, to examine L2 learners' cognitive and behavioral engagement with automated feedback provided by Write & Improve. Specifically, it investigated how the explicitness and accuracy of feedback influence learners' engagement during the writing process. The findings indicate that participants allocated significantly more time and cognitive resources to processing indirect feedback compared to direct feedback. However, the uptake rate of indirect feedback was lower, and the revisions made in response to such feedback were less effective in improving written output.

Despite its effectiveness in reducing errors and enhancing accuracy, research indicates that AWE has limited influence on other dimensions of writing. Kellogg et al. (2010) found that Criterion enhanced mechanics but had little effect on writing quality. Similarly, Chen and Cheng (2008) investigated the effectiveness of MyAccess and reported that many students

perceived the feedback as unclear and mechanical, leading them to believe their writing had not improved, suggesting AWE may be more beneficial for beginner or intermediate learners. Furthermore, the timing of AWE feedback—whether immediate or delayed—has been shown to significantly impact writing accuracy (Lavolette, 2015). Li et al. (2017) examined the short- and long-term effects of Criterion feedback on grammatical accuracy, finding that AWE had a stronger short-term impact, leading to a significant reduction in errors across eight categories. However, in the long term, significant improvement was observed in only one category.

A key factor influencing the effectiveness of AWE programs is students' willingness to adopt and engage with the technology (Li et al., 2019). Merely providing feedback does not guarantee its effective use. For instance, Chappelle et al. (2015) found that undergraduate students ignored feedback from Criterion more than half the time. However, those who actively engaged with the feedback demonstrated significant improvements in their writing. Research also indicates that AWE can enhance teacher feedback quality. Wilson and Czik (2016) found that teachers using PGE focused more on higher-level writing skills, such as organization and idea development, than those without AWE. This suggests that AWE can complement traditional feedback, addressing grammar and mechanics while teachers emphasise content and ideas.

Despite extensive research on the impact of AWE on writing development, relatively few studies have investigated its effectiveness in test performance and standardised language assessments (Shermis, 2014), especially given the fact that several AWE programs were developed as scoring systems for standardised tests (Shermis 2014). While MyAccess and Criterion have been widely studied, only two empirical studies have examined Write & Improve for writing development and feedback (Karpova, 2020; Liu & Yu, 2022). With assessment shifting toward learning-oriented approaches (Jones & Saville, 2016), it is crucial to explore how AWE can support such approaches and influence writing development. This study addresses these gaps by investigating the effectiveness of Write & Improve in enhancing IELTS writing performance among Thai students preparing for the exam.

Methodology

Research Contexts and Participants

The participants were English major students enrolled in the "English Preparation for Standardized Tests" course at a public university in Thailand. Initially, all 55 students in the class were invited to participate; however, two withdrew in Week 9, leaving a final sample of 53 participants at the study's

conclusion. Questionnaire data revealed that the sample consisted of 30 female and 23 male participants, aged between 20 and 23 years. The majority had been learning English for over 10 years. Notably, more than half (N = 31) had prior experience with AWE programs before the intervention, while 22 participants had never used such programs. Regarding English proficiency, participants were classified based on the scores provided by the university language proficiency test. Eighteen students were at B1, 29 at B2, and 6 at C1 level.

Table 1

Demographic Information of the Participants (1)

Age	F	%	Gender	F	%	English Language Proficiency	F	%
20	27	50.95	Male	23	43.40	A2	0	0.00
21	19	35.85	Female	30	56.60	B1	18	12.05
22	6	11.3				B2	29	63.65
23	1	1.90				C1	6	24.25

Table 2

Demographic Information of the Participants (2)

Experience using AWE programs	F	%	Years of learning English	F	%
Yes	31	58.50	10-15	17	32.10
No	22	41.50	15-20	31	58.50
			20 or more	5	9.40

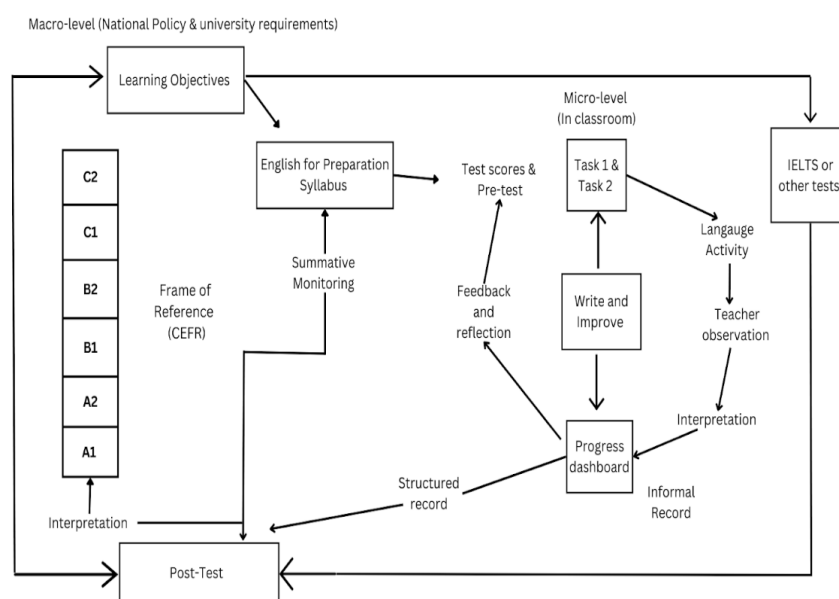
The "English Preparation for Standardized Tests" course is an elective for English major students at the university, running for 16 weeks with three hours of instruction per week. The course focuses on essential skills for standardized language tests, historically alternating between IELTS and TOEFL based on student needs. However, IELTS has been the primary focus for the past five years. The curriculum covers all four IELTS skills: reading (Weeks 1–3), writing (Weeks 4–8), listening (Weeks 9–11), and speaking (Weeks 12–14). Week 15 is dedicated to course reflection, followed by the final examination in Week 16.

Learning-oriented Assessment Design

Write and Improve was integrated into *English Preparation for Standardised Tests* with an embedded framework of the systemic model of LOA (Jones & Saville, 2016). At the macro level, the university requires CEFR B2 proficiency for graduation in line with national expectations (B1–B2). Accordingly, the course was designed to prepare students for standardised tests. At the micro level, a medium-cycle LOA approach was implemented to support the development of students' writing skills (William, 2006). In this course, Write and Improve was used as an *LOA activity*. Both task 1 and task 2 were used as *Task* that create language activity, involving writing, receiving AWE feedback, making revisions and resubmitting the drafts. The Class View of Write and Improve enabled the teacher to observe and interpret the students' writing performance and monitor their learning progress through a learning dashboard. The structured record is also maintained through the CEFR scores assigned to the students' writing. The students also took a post-test after they had finished the LOA activity, which can be seen as a record of achievement. After they have finished this course, they may go on to take IELTS or other local and international standardised tests to prove their language proficiency to the university. The LOA design in this study is displayed in Figure 1 below:

Figure 1

Write and Improve and LOA design (Adapted from Jones & Saville, 2016)



Write and Improve and the Intervention

Write & Improve (<https://writeandimprove.com/>) is a free AWE platform that evaluates writing proficiency and provides instant feedback aligned with the Common European Framework of Reference for Languages (CEFR). The system assesses four key aspects: content, communicative achievement, organization, and language (Thao & Vu, 2023). The program offers three main features: W&I Workbook, Test Zone, and Class View. In the W&I Workbook, students can select from five categories, including three proficiency levels (beginner, intermediate, advanced), as well as W&I business and W&I just for fun. Upon submission, the system evaluates the writing according to the CEFR, providing instant feedback at the word and sentence levels. The feedback is visually enhanced with distinct colours and symbols (see Fig. 1).

Figure 2

Feedback from Write and Improve

& *That's amazing. Your writing is really improving. Pay attention to the feedback. What changes can you make? Why not try to write longer sentences? Keep writing to keep improving.*

The diagram is about making drinking glass showing step by step using visual and label to describe a 9 step process, starting with melting the glass, pressing the hot glass, shaping it, firing in a kiln, cooling it, finally inspecting it to decide if it is a waste or good product. **to explained it more,** Step one : throwing the glass into the furnace with 105 c temperature. **From there the glass is put into the mould and pressed into container which is have thick wall,** **it** presses into a second mould and **air is blown to form a glass.** Then the glass is **placed** in the kiln at **after that the worker inspected** decided if it was a waste or a **and bad one.** [feedback' page](#)

☆ Suspicious word X

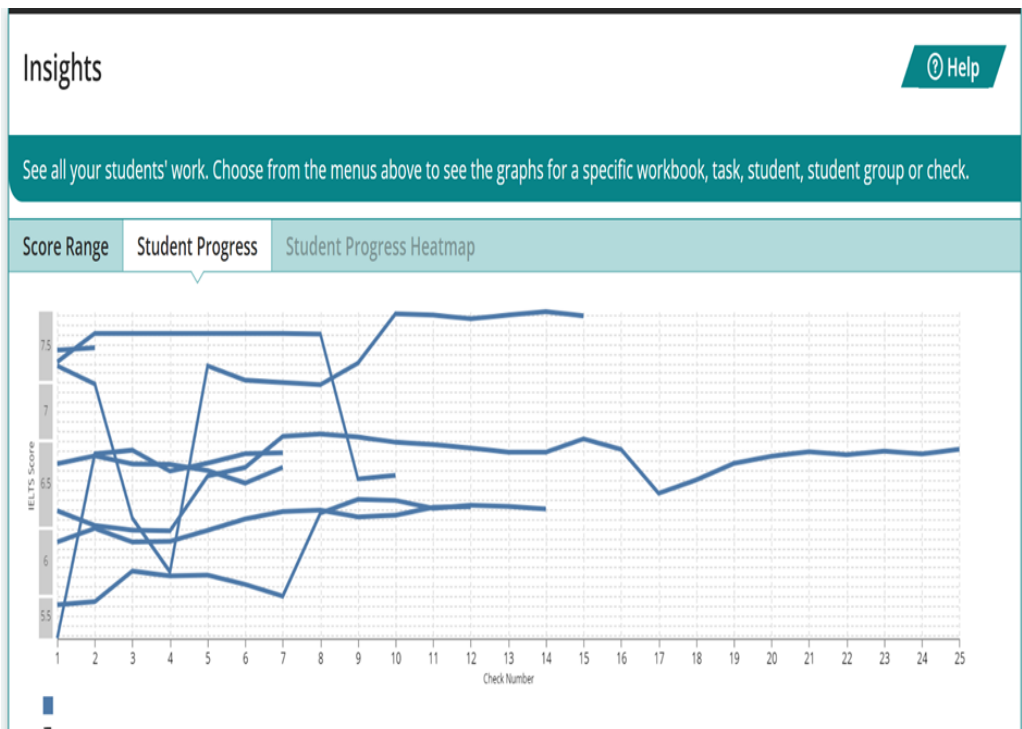
Is this the correct word?
This doesn't look right to us

Next

In the Test Zone, students can practice writing for both Academic and General Training IELTS. Submitted drafts receive feedback along with band scores for all four IELTS criteria. The Class View feature enables teachers to create writing classes, design tasks, and customize workbooks, while also providing graphs and insights into student and class progress (see Fig. 2).

Figure 3

A Student's Progress from Class View



In this study, Write & Improve served as a learning intervention. Over five weeks, 53 participants used the platform for self-practice. Before the intervention, the participants completed the pre-test, and an introductory session guided them on program usage, feedback interpretation (colors and symbols), and the intervention timeline was conducted. The participants were instructed not to use other AWE tools or writing assistants during the intervention. The researcher conducted an introductory session in the final two hours of Week 8, outlining the requirements. They were asked to complete one Task 1 and one Task 2 question per week but could submit additional tasks. Participants were also free to revise and resubmit their drafts

as often as wanted. The four-week intervention ran from Weeks 8 to 12. At the end of week 12, the participants completed the post-test. The intervention is summarised in Table 3 below:

Table 3

Write and Improve Intervention

Week	Actions by learners	Actions by teachers
Week 8	• Pre-tests • Attended Introductory Sessions	• Conducted Introductory session • Ran pre-tests
Week 9	• Writing Practice with AWE	• Monitored Learning Progress in the dashboard
Week 10	• Writing Practice with AWE	• Monitored Learning Progress in the dashboard
Week 11	• Writing Practice with AWE	• Monitored Learning Progress in the dashboard
Week 12	• Writing Practice with AWE • Post-tests	• Monitored Learning Progress in the dashboard • Ran post-tests

Data Collection and Analysis

A pre-test (See appendix) was conducted in Week 9 of the course, followed by a post-test (See appendix) in Week 13, both administered face-to-face in the classroom. The researcher personally proctored both assessments. Participants completed two writing tasks (Task 1 and Task 2) within 60 minutes. Their responses were assessed by two current IELTS examiners. Both examiners marked the tests independently using the IELTS writing band descriptors. Also, Cohen's Kappa analysis was performed using SPSS. The analysis yielded a value of .956, indicating acceptable inter-rater reliability. Scores for each criterion—task achievement/response, lexical resource, grammatical range and accuracy, and coherence and cohesion—were recorded in Excel spreadsheets for further analysis. The scores were analysed using SPSS 28.0 (Statistical Package for the Social Sciences). Paired sample t-tests were conducted to compare scores across each criterion, identifying differences between pre-test and post-test performance.

Learning logs from 53 participants (August–September 2023) were collected via the Classview system in Write & Improve. These logs recorded the number of completed tasks and the frequency of writing checks (instances where participants submitted their writing for AWE feedback). As the data

were system-generated, they were already quantified for statistical analysis. To examine the relationship between learning behaviors and writing improvement, a binary logistic regression model was developed and applied.

Ethical Considerations

Ethical approval for the study was obtained from the University ethics committee prior to the intervention. Participants received a consent form and a participant information sheet before engaging with Write & Improve. They were informed that participation was voluntary and that they could withdraw at any stage, even after the study had commenced. All collected data were accessible only to the researchers of the current study. While participants were initially identifiable, their identities were anonymized through the assignment of pseudonyms.

Findings

The Effects of Write and Improve on Students' Writing Performance

To examine score differences between pre-test and post-test performance, paired-sample t-tests were conducted across eight criteria: task achievement, lexical resource, grammatical range and accuracy, and coherence and cohesion for Task 1, and task response, lexical resource, grammatical range and accuracy, and coherence and cohesion for Task 2.

Differences between Pre- and Post-Test Scores on Writing Task 1

Table 4

Results from a Paired-Sample t-Test on the Scores for Task Achievement

Test	N	M	SD	T	Df	Sig. (p)
Pre-test	53	5.92	0.92	-2.53	52	.014
Post-test	53	6.23	0.78			

Note: N = number of subjects, M = mean of scores, SD = Standard deviation t = t scores, Df = degrees of freedom Sig. = Significance.

As shown in Table 4, participants demonstrated significantly higher Task Achievement scores in the post-test (M = 6.23, SD = 0.73) compared to the pre-test (M = 5.92, SD = 0.87), with a statistically significant difference at the 0.05 level ($t(52) = -2.53, p = .014$). These findings indicate that

participation in Write & Improve was associated with improved Task Achievement performance in the post-test.

Table 5

Results from a Paired-Sample t-Test on the Scores for Lexical Resources

Test	N	M	SD	T	Df	Sig. (p)
Pre-test	53	6.28	0.72	-3.21	52	.002
Post-test	53	6.60	0.53			

Note: N = number of subjects, M = mean of scores, SD = Standard deviation t = t scores, Df = degrees of freedom Sig. = Significance.

As indicated in Table 5, participants achieved significantly higher Lexical Resource scores in the post-test (M = 6.60, SD = 0.53) compared to the pre-test (M = 6.28, SD = 0.72), with a statistically significant difference at the 0.05 level ($t(52) = -3.21, p = .002$). These results suggest that following the Write & Improve intervention, participants demonstrated improved lexical resource performance in the post-test performance in the post-test.

Table 6

Results from a Paired-Sample t-Test on the Scores for Grammatical Range and Accuracy (Task 1)

Test	N	M	SD	T	Df	Sig. (p)
Pre-test	53	6.02	0.72	-7.78	52	.001
Post-test	53	6.79	0.57			

Note: N = number of subjects, M = mean of scores, SD = Standard deviation t = t scores, Df = degrees of freedom Sig. = Significance.

It can be seen that the participants scored significantly higher in Grammatical Range and Accuracy in the post-test (M = 6.79, SD = 0.57) compared to the pre-test (M = 6.02, SD = 0.72), with a statistically significant difference at the 0.05 level ($t(52) = -7.78, p = .001$). These findings suggest that engagement with Write & Improve contributed to improved grammatical accuracy and range of sentence structure used in the post-test.

Table 7

Results from a Paired-Sample t-Test on the Scores for Coherence and Cohesion

Test	N	M	SD	T	Df	Sig. (p)
Pre-test	53	5.98	0.89	-5.52	52	.001
Post-test	53	6.55	0.61			

Note: N = number of subjects, M = mean of scores, SD = Standard deviation t = t scores, Df = degrees of freedom Sig. = Significance.

As indicated, the participants achieved significantly higher Coherence and Cohesion scores in the post-test ($M = 6.55$, $SD = 0.61$) compared to the pre-test ($M = 5.98$, $SD = 0.89$), with a statistically significant difference at the 0.05 level ($t(52) = -5.52$, $p = .001$). This means that engagement with Write & Improve contributed to better performance in the post-test as regards coherence and cohesion. In summary, the Write & Improve intervention had a positive impact on participants' Task 1 writing performance across all four criteria. Notably, the most substantial improvement was observed in Grammatical Range and Accuracy.

Differences between Pre and Post-Test Scores on Writing Task 2

Table 8

Results from a Paired-Sample t-Test on the Scores for Task Response

Test	N	M	SD	T	Df	Sig. (p)
Pre-test	53	6.32	0.78	-1.99	52	.052
Post-test	53	6.57	0.87			

Note: N = number of subjects, M = mean of scores, SD = Standard deviation t = t scores, Df = degrees of freedom Sig. = Significance.

Despite the participants achieving higher Task Response scores in the post-test ($M = 6.57$, $SD = 0.87$) compared to the pre-test ($M = 6.32$, $SD = 0.78$), the difference was not statistically significant at the 0.05 level ($t(52) = -1.99$, $p = .052$). This suggests that while participants showed some improvement after engaging with Write & Improve, the gains in Task Response were not statistically significant.

Table 9*Results from a Paired-Sample t-Test on the Scores for Lexical Resources*

Test	N	M	SD	T	Df	Sig. (p)
Pre-test	53	6.42	0.63	-4.19	52	.001
Post-test	53	6.77	0.47			

Note: N = number of subjects, M = mean of scores, SD = Standard deviation t = t scores, Df = degrees of freedom Sig. = Significance.

As shown in Table 9, the participants scored higher in Lexical Resources in the post-test (M = 6.77, SD = 0.47) compared to the pre-test (M = 6.42, SD = 0.63), with a statistically significant difference at the 0.05 level ($t(52) = -4.19, p = .001$). As a result, it can be assumed that after engaging with Write and Improve, the participants performed better in terms of lexical resources in the post-test.

Table 10*Results from a Paired-Sample t-Test on the Scores for Grammatical Range and Accuracy (Task 2)*

Test	N	M	SD	T	Df	Sig. (p)
Pre-test	53	6.13	0.71	-6.21	52	.001
Post-test	53	6.85	0.77			

Note: N = number of subjects, M = mean of scores, SD = Standard deviation t = t scores, Df = degrees of freedom Sig. = Significance.

The participants demonstrated significantly higher Grammatical Range and Accuracy scores in the post-test (M = 6.85, SD = 0.77) compared to the pre-test (M = 6.13, SD = 0.71), with a statistically significant difference at the 0.05 level ($t(52) = -6.21, p = .001$). These results indicate that after five weeks of using Write & Improve, participants showed notable improvements in grammatical range and accuracy.

Table 11*Results from a Paired-Sample t-Test on the Scores for Coherence and Cohesion (Task 2)*

Test	N	M	SD	T	Df	Sig. (p)
Pre-test	53	6.08	0.70	-4.95	52	.001
Post-test	53	6.58	0.60			

Note: N = number of subjects, M = mean of scores, SD = Standard deviation t = t scores, Df = degrees of freedom Sig. = Significance.

As depicted in Table 11, the scores in the post-test ($M = 6.58$, $SD = 0.60$) were much higher than those in the pre-test ($M = 6.08$, $SD = 0.70$), with a significant difference at a 0.05 probability level ($t(52) = -4.95$, $p = .001$). These findings suggest that engagement with Write & Improve contributed to improved coherence and cohesion in participants' writing performance.

In conclusion, the *Write & Improve* intervention had a positive effect on participants' Task 2 writing across all four criteria. However, statistically significant improvements were observed in three criteria, with Task Response being the exception. Similar to Task 1, the most substantial gains were recorded in Grammatical Range and Accuracy.

The Effects of AWE on Different Groups of Students

To examine AWE's impact across proficiency levels, students were categorised into three groups based on pre-test scores: intermediate (≤ 5.5), upper-intermediate (6.0–6.5), and advanced (≥ 7.0). Descriptive statistics were used to assess the number of students in each group who showed score improvements.

Table 12

Score Improvements by Different Groups of Students (Task 1)

Group	N (53)	Increased	Unchanged	Decreased
5.5 and below	10	9 (90%)	1 (10%)	-
6.0–6.5	29	20 (69%)	5 (17.2%)	4 (13.8 %)
7.0 and above	14	5 (35.7%)	7 (50%)	2 (14.3%)

As shown in Table 12, 90% of students in the ≤ 5.5 group demonstrated score gains in the post-test, while nearly 70% of those in the 6.0–6.5 group also improved. In contrast, only about one-third of students in the ≥ 7.0 group saw improvements, with over 60% maintaining or receiving lower scores. These results suggest that Write & Improve was most beneficial for students at the ≤ 5.5 proficiency level in Task 1.

Table 13*Score Improvements by Different Groups of Students (Task 2)*

Group	N (53)	Increased	Unchanged	Decreased
5.5 and below	15	14 (93.3%)	-	1 (6.7%)
6.0-6.5	27	18 (66.7)	8 (29.6%)	1 (3.7%)
7.0 and above	11	1 (9.1%)	8 (72.7%)	2 (18.2%)

As shown in Table 13, 93.3% of students in the ≤ 5.5 group improved in the post-test after using Write & Improve for five weeks. Similarly, about two-thirds of students in the 6.0–6.5 group showed score gains. However, in the ≥ 7.0 group, only one student (9.1%) improved, while most received the same or lower scores. These findings suggest that while Write & Improve benefited intermediate learners (≤ 5.5), its impact on advanced students (≥ 7.0) was minimal.

The Relationship between Learning Behaviours and Writing Improvements

A logistic regression model was used to examine the relationship between learning behaviors and writing improvement, defined as score gains in the post-test. The analysis identified factors statistically associated with writing progress. The model was based on two AWE-generated learning metrics: (1) the number of completed tasks and (2) the number of times students had their writing checked by Write & Improve. In total, 53 students completed 406 Task 1 and 406 Task 2 responses, with their writing checked 1,586 times by the program.

Table 14*Logistic Regression Analysis of the Learning Behaviours in Task 1*

	B	S.E.	Wald	df	Sig.	Exp(B)
Task Achievement						
Number of tasks	0.463	0.232	3.997	1	0.046*	1.589
Number of checks	0.007	0.015	0.235	1	0.628	0.993
Constant	-3.806	1.701	5.006	1	0.025	0.022
Lexical Resources						
Number of tasks	0.534	0.265	4.056	1	0.044*	1.706

Number of checks	0.01	0.019	0.272	1	0.602	1.01
Constant	-5.491	2.308	7.257	1	0.007	0.004
Grammatical Range and Accuracy						
Number of tasks	1.064	0.355	8.975	1	0.003*	2.898
Number of checks	0.028	0.018	2.318	1	0.128	0.972
Constant	6.669	2.435	7.503	1	0.006	0,001
Coherence and Cohesion						
Number of tasks	0.749	0.286	6.859	1	0.009*	2.115
Number of checks	-0.012	0.017	0.549	1	0.459	0.988
Constant	-5.417	2.059	6.992	1	0.009	0.004

As shown in Table 14, the number of completed tasks was statistically significant ($p = 0.05$) in predicting writing improvement across all four Task 1 criteria. This indicates that students who completed more tasks were more likely to achieve higher post-test scores. However, the frequency of draft checks had no statistically significant effect on writing improvement for any criterion.

The relationship between learning behaviours and writing improvement in Task 2 was also analysed. The results are presented in Table 15 below.

Table 15

Logistic Regression Analysis of the Learning Behaviours in Task 2

	B	S.E.	Wald	df	Sig.	Exp(B)
Task Achievement						
Number of tasks	-0.157	0.18	0.766	1	0.382	0.854
Number of checks	0.016	0.015	1.155	1	0.282	1.061
Constant	0.344	1.261	0.074	1	0.785	1.411
Lexical Resources						
Number of tasks	0.401	0.234	2.937	1	0.087	1.494
Number of checks	0.01	0.019	0.31	1	0.578	1.01
Constant	-3.83	1.72	4.96	1	0.026	0.022
Grammatical Range and Accuracy						

Number of tasks	-0.31	0.223	1.931	1	0.165	0.733
Number of checks	0.064	0.033	3.652	1	0.051	1.066
Constant	1.24	1.387	0.8	1	0.371	3.456

Coherence and Cohesion

Number of tasks	0.081	0.186	0.188	1	0.664	1.084
Number of checks	0.018	0.019	0.838	1	0.36	1.018
Constant	-1.05	1.286	0.666	1	0.464	0.35

As indicated in Table 14, the frequency of writing checks was statistically significant in relation to improvements in grammatical range and accuracy but not for other Task 2 criteria. This suggests that students who reviewed their drafts more frequently performed better in grammar on the post-test. Notably, unlike Task 1, the number of completed tasks had no statistically significant impact on writing performance across any Task 2 criteria.

Discussion

Effects of Write and Improve on Students' IELTS Writing

The post-test score gains across all four criteria in both Task 1 and Task 2 suggest that Write & Improve had a positive impact on students' writing performance. For Task 1 writing, statistical analysis revealed significant improvements across all four criteria, with the most substantial gains observed in Grammatical Range and Accuracy ($M = 0.77$) and Coherence and Cohesion ($M = 0.57$). Although score gains for Task Achievement and Lexical Resources were less pronounced, they remained statistically significant. For Task 2 writing, statistical analysis showed significant score improvements, though not across all four criteria. While Task Response showed a 0.25-point gain, this improvement was not statistically significant. Consistent with Task 1, the most substantial gains were observed in Grammatical Range and Accuracy ($M = 0.72$) and Coherence and Cohesion ($M = 0.50$). Empirically, since students did not receive direct writing instruction during the intervention, it is reasonable to infer that the positive changes in post-test performance were associated with, though not entirely due to, the Write & Improve intervention.

These findings yield several key observations. Firstly, the most substantial score gains were observed in grammatical range and accuracy. This could be because, like many other AWE programs, Write and Improve (in its current iteration) is most effective in providing grammar and punctuation-related feedback. Such results support previous research indicating that AWE

enhances writing accuracy (El Ebyary & Windeatt, 2010; Li et al., 2015; Wang et al., 2013). From a theoretical perspective, written corrective feedback (WCF), including that generated through automated writing evaluation (AWE) systems, can facilitate learners' ability to notice, interpret, and self-correct linguistic errors on the basis of the feedback provided. Such engagement with WCF is likely to direct learners' attention more explicitly towards underlying grammatical forms and rules, thereby fostering heightened grammatical awareness and accuracy. Secondly, the relatively smaller score gains in Task Achievement (Task 1) and Task Response (Task 2), with the latter not reaching statistical significance, were expected. Given the nature of AWE feedback, how well writing responds to a prompt is typically not emphasised. This type of evaluation likely requires a more human-centered approach, as opposed to the mechanistic feedback provided by AWE systems. Again, this result is in line with previous studies finding that AWE has limited effects on writing beyond grammatical accuracy (Chen & Cheng, 2008; Kellogg et al., 2010).

An analysis of AWE's impact across proficiency levels revealed that, for both Task 1 and Task 2, the program was most effective in enhancing overall scores among intermediate learners (≤ 5.5), followed by those in the upper-intermediate group (6.0–6.5). It is interesting, however, to find that Write and Improve only showed a moderately positive impact on task 1 for the advanced group (7.0 and above) and almost no effects at all for task 2. This outcome may be attributed to the nature of AWE feedback, which primarily targets accuracy, mechanics, and lexical suggestions, areas particularly relevant to intermediate-level learners. In contrast, students at higher proficiency levels may have already mastered these linguistic aspects and require more specific feedback on content development and stylistic features of their writing. These findings align with existing AWE research suggesting that AWE is most suitable for beginner and intermediate learners (Chen & Cheng, 2008; Lavolette, 2015). It is also possible that there is a sweet spot in the middle of the proficiency range where progress is possible over a short period.

Learning Behaviours and Improvements

For Task 1 writing, logistic regression analysis indicated that students who completed more writing tasks were significantly more likely to achieve higher post-test scores across all four criteria. However, the frequency of writing checks did not show a statistically significant correlation with score improvements. For Task 2, logistic regression analysis revealed no statistically significant correlation between the number of completed tasks or the frequency of writing checks and post-test score improvements. Several key

observations emerge from the findings. For Task 1, it is unsurprising that students who completed more tasks were more likely to achieve higher post-test scores. Completing more tasks required students to produce multiple drafts, construct varied sentence structures, select appropriate vocabulary, and analyse new data presented in graphs. This process likely enhanced their writing competence and improved their ability to approach writing tasks effectively.

The lack of correlation between the number of writing checks and score improvements, however, surprised us. One possible explanation is that, although students frequently reviewed and revised their work, the modifications made in each revision were likely incremental, resulting in limited overall impact on their writing scores. For instance, students may have made only minor revisions—such as modifying one or two aspects of their writing—before resubmitting their drafts for AWE feedback, repeating this process across multiple revisions. If this was the case, it is understandable that a higher number of checks did not necessarily correlate with better post-test performance. This suggests that in AWE-based revision, the quality of each revision, rather than the quantity, is likely to have a greater impact on writing development. For Task 2, unlike Task 1, which follows more structured writing patterns and prescriptive grammatical structures, students are required to demonstrate greater creativity and productivity in both idea development and sentence construction. This inherent variability likely contributed to differences in performance among students. Additionally, Task 1 exhibited greater score gains across all four criteria in the post-test compared to Task 2, which may explain the lack of statistically significant correlations between learning behaviors and writing improvements in Task 2.

Conclusion and Implications

While the findings of this study are encouraging, certain methodological and pedagogical limitations should be acknowledged. From a methodological perspective, although students were instructed to refrain from using other writing assistant programs during the intervention, it is possible that they may have intentionally or inadvertently received additional instructional support from external sources. Therefore, it cannot be conclusively determined that the observed score improvements were solely attributable to the Write & Improve intervention. Additionally, due to administrative constraints, the study lacked a control group, preventing direct comparison between students using Write & Improve and those following a business-as-usual approach. To enhance the robustness of future research, a quasi-experimental design incorporating both experimental and control groups is recommended. Furthermore, as this study was small-scale (N = 53)

and conducted with a Thai cohort, the findings may not be generalizable to broader populations or diverse contexts. Therefore, future research should consider a larger-scale study in a different setting to enhance external validity. Finally, as this study spanned only five weeks, the duration may have been insufficient for substantial language improvement, especially for the low proficiency students. Therefore, future research should adopt a longitudinal design extending over a semester to better assess the long-term effects of AWE interventions.

The findings of this study suggest that Write & Improve, along with other AWE programs, has become relatively effective in identifying grammatical errors, with feedback proving beneficial for enhancing grammar and writing accuracy. However, as AWE remains limited in providing feedback beyond grammar and mechanics, a hybrid feedback model that integrates automated and human feedback is likely essential for comprehensive writing development. In this regard, AWE can serve as an initial feedback tool, primarily addressing grammar and mechanics, while teachers focus on providing feedback related to content and organisation. Implementing a hybrid feedback model not only alleviates teachers' workload but also enhances both their efficiency and the overall quality of feedback delivered to students.

Moreover, the results of this study also suggest that when using AWE as a writing practice tool, teachers may achieve better outcomes by encouraging students to engage with multiple tasks rather than making incremental revisions to a single task. This approach is particularly relevant if AWE is utilized as a self-study tool, maximising its effectiveness in writing development. Also, learners can be encouraged to choose topics they are interested in to encourage better engagement with the task and to stimulate higher motivation to persevere in practicing. From a methodological standpoint, learning analytics offers a valuable approach to examining the complex factors influencing learning outcomes. If technological capabilities allow, future research could incorporate additional learning data, such as student profiles, login frequency, and time spent on tasks and revisions, to provide a more comprehensive understanding of the relationship between learning behaviors and writing development.

About the Authors

Napat Jitpaisarnwattana: A lecturer of English and Computer-assisted Language Learning at Silpakorn University, Thailand. He has recently finished a Master of Studies focusing on Machine Learning and Automated Language Assessment at Homerton College, University of Cambridge. He is an associate editor of *rEFlections Journal* and an associate editor of the

International Journal of Southeast Asian Media Studies. His research interests include LMOOCs, the Internet of things, digital wellbeing, digital literacies, learning analytics, machine learning, natural language processing and AI in language education.

Nick Saville: Director of Research and Thought Leadership, Cambridge Assessment English. His research interests are language proficiency scales and descriptors in the context of assessment (e.g. the CEFR and English Profile Programmed); learning-oriented assessment, including socio-cognitive models of learning and assessment, pedagogical grammars of English and interfaces with SLA; language and migration, including multilingualism and the impact of language assessment in social contexts.

References

- Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Erlbaum.
- Chappelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3), 385–405. <https://doi.org/10.1177/0265532214565386>
- Chen, C.-F. E., & Cheng, W.-Y. E. C. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, 12(2), 94–112.
- Chou, H. N., & Chung, C. M. (2013). EFL college students' writing performance and perceptions of the use of an automated writing evaluation system. *Journal of English Education*, 2(1), 93–126.
- Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement*. New York, NY: Palgrave Macmillan.
- Cotos, E. (2018). Automated writing evaluation. In J. I. Lontas (Ed.), *The TESOL Encyclopedia of English Language Teaching*. Wiley. (pp. 1-6).
- Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1), 1–36.
- El Ebyary, K., & Windeatt, S. (2010). The impact of computer-based feedback on students' written work. *International Journal of English Studies*, 10(2), 121–142.
- Fan, N. (2023). Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study.

- SAGE Open*, 13(2).
<https://doi.org/10.1177/21582440231181296>
- Ferris, D. R., & Hedgcock, J. S. (2005). *Teaching ESL composition: Purpose, process, and practice* (2nd ed.). Lawrence Erlbaum Associates.
- Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, 63, 480–499.
- Hockly, N. (2019). Automated writing evaluation. *ELT Journal*, 73(1), 82–88.
<https://doi.org/10.1093/elt/ccy044>
- Jitpaisarnwattana, N. (2024). Idea sharing: Using ChatGPT for diagnostic writing assessment. *PASAA*, 69, 561–574.
- Jitpaisarnwattana, N., Darasawang, P., & Reinders, H. (2022). Delving into personalisation behaviours in a language MOOC. *International Journal of Learning, Teaching and Educational Research*, 21(1), 92–108.
doi:10.26803/ijlter.21.1.6
- Jitpaisarnwattana, N., & Saville, N. (2025). Technology and Learning-Oriented Assessment (LOA). In L. Lontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-6).
<https://doi.org/10.1002/9781118784235.eelt1092>
- Jitpaisarnwattana, N., & Saville, N. (2025). Autonomous learning and students' perceptions of automated writing evaluation as a tool to improve writing skills . *Technology in Language Teaching & Learning*, 7(3), 102826. <https://doi.org/10.29140/ttl.v7n3.102826>
- Jones, N., & Saville, N. (2016). Learning oriented assessment: A systemic approach. *Studies in Language Testing* 45. Cambridge University Press.
- Karpova, K. (2020). Integration of "Write and Improve" AWE tool into EFL a higher educational establishment: case study. *Celtic: A Journal of Culture, English Language Teaching, Literature and Linguistics*, 7(2), 137–150. <https://doi.org/10.22219/celtic.v7i2.14036>
- Kellogg, R. T., Whiteford, A. P., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research*, 42, 173–196.
- Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language, Learning & Technology*, 19(2), 50–68.
- Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C., & Xiao, W. (2019). Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Computer Assisted Language Learning*, 32(7), 784–804. <https://doi.org/10.1080/09588221.2018.1540433>
- Li, Z., Feng, H.-H., & Saricaoglu, A. (2017). The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *Calico Journal*. 34, 355–375.

- Liao, H. C. (2016). Enhancing the grammatical accuracy of EFL writing by using an AWE-assisted process approach. *System*, 62, 77–92.
- Liu, S., & Yu, G. (2022). L2 learners' engagement with automated feedback: An eye tracking study. *Language Learning & Technology*, 26(2), 78–105. <https://doi.org/10125/73480>
- Parra, L., & Calero, X. (2019). Automated writing evaluation tools in the improvement of the writing skill. *International Journal of Instruction*, 12(2), 209–226. <https://doi.org/10.29333/iji.2019.12214a>
- Ranalli, J., & Hegelheimer, V. (2022). Introduction to the special issue on automated writing evaluation. *Language Learning & Technology*, 26(2), 1–4. <http://doi.org/10125/73473>
- Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1), 8–25. doi:10.1080/01443410.2015.1136407
- Rezaei, A. R., & Lovorn, M. (2010) Reliability and validity of rubrics for assessment through writing. *Assessing Writing* 15 (1) 18–39.
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76. doi:10.1016/j.asw.2013.04.001
- Shi, Z., Liu, F., Lai, C., & Jin, T. (2022). Enhancing the use of evidence in argumentative writing through collaborative processing of content-based automated writing evaluation feedback. *Language Learning & Technology*, 26(2), 106–128. <https://doi.org/10125/73481>
- Tang, J., & Wu, Y. A. (2017). Impact of AWE rubrics and automated assessment on EFL writing instruction. *International Journal of Computer-Assisted Language Learning and Teaching*, 7, 58–74.
- Thao, N., & Vu, T., (2023). Using write & improve to improve EFL learners' writing skills. *Computer Assisted Language Learning Electronic Journal (CALL-EJ)*, 24(1), 198–204.
- Ngo, N., Chen, J., & Lai, W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments*, 32(2), 727–744. <https://doi.org/10.1080/10494820.2022.2096642>
- Wang, P. L. (2015). Effects of an automated writing evaluation program: Student experiences and perceptions. *Electronic Journal of Foreign Language Teaching*, 12(1), 79–100.
- Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language

- university students' writing. *Computer Assisted Language Learning*, 26(3), 234–257. <https://doi.org/10.1080/09588221.2012.655300>
- Ware, P. D., & Warschauer, M. (2006). Electronic feedback and second language writing. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Context and issues* (pp. 104–122). Cambridge University Press.
- Warschauer, M., & Grimes, G. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, 3, 22–36.
- Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94–109. <https://doi.org/10.1016/j.compedu.2016.05.004>
- Woodworth, J., & Barkaoui, K. (2020). Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal*, 37(2), 234–247.
- Yannakoudakis, H., Andersen, Ø. E., Geranpayeh, A., Briscoe, T., & Nicholls, D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31, 251–267.
- Zhai, N., & Ma, X. (2022) Automated writing evaluation (AWE) feedback: A Systematic investigation of college students' acceptance, *Computer Assisted Language Learning*, 35(9), 2817-2842.
doi:10.1080/09588221.2021.1897019
- Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning* 1(1), 170–176.
<https://doi.org/10.1515/jccall-2021-2007>

Appendix

Pre-and Post-tests

Pre-test: Task 1

WRITING

WRITING TASK 1

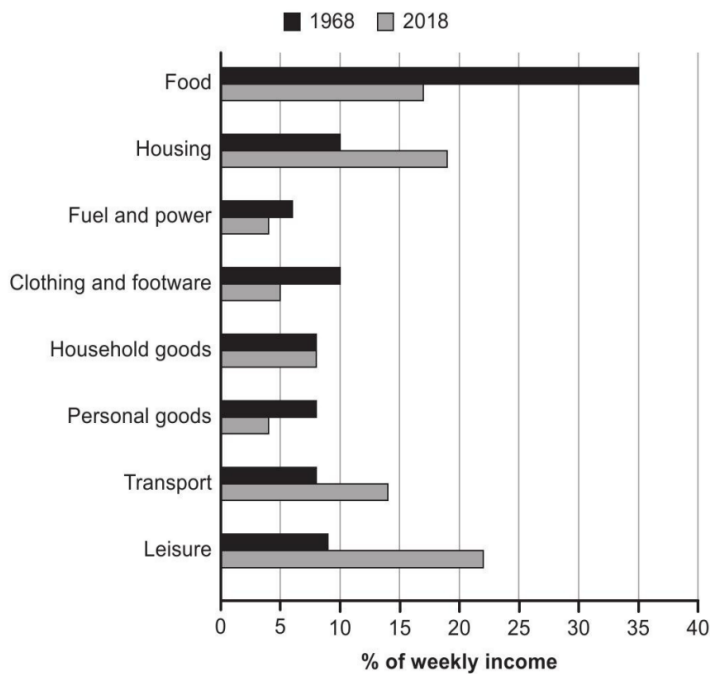
You should spend about 20 minutes on this task.

The chart below gives information about how families in one country spent their weekly income in 1968 and in 2018.

Summarise the information by selecting and reporting the main features, and make comparisons where relevant.

Write at least 150 words.

1968 and 2018: average weekly spending by families



Pre-test: Task 2

WRITING TASK 2

You should spend about 40 minutes on this task.

Write about the following topic:

Some people say that advertising is extremely successful at persuading us to buy things. Other people think that advertising is so common that we no longer pay attention to it.

Discuss both these views and give your own opinion.

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Write at least 250 words.

Post-test: Task 1

Test 1

WRITING

WRITING TASK 1

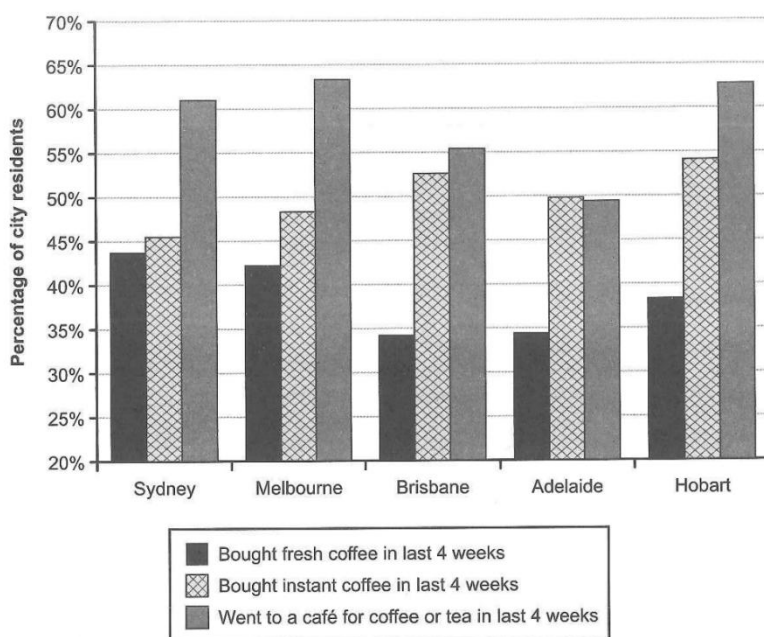
You should spend about 20 minutes on this task.

The chart below shows the results of a survey about people's coffee and tea buying and drinking habits in five Australian cities.

Summarise the information by selecting and reporting the main features, and make comparisons where relevant.

Write at least 150 words.

Coffee and tea buying and drinking habits in five cities in Australia



Post-test: Task 2

WRITING TASK 2

You should spend about 40 minutes on this task.

Write about the following topic:

Some people say that advertising is extremely successful at persuading us to buy things. Other people think that advertising is so common that we no longer pay attention to it.

Discuss both these views and give your own opinion.

Give reasons for your answer and include any relevant examples from your own knowledge or experience.

Write at least 250 words.