Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327

Arab World English Journal (AWEJ) Special Issue on CALL Vol. 17, No.12. July 2026 Pp. 237-256
DOI: https://dx.doi.org/10.24093/awej/call12.14

Evaluating the Automated Scoring and Feedback provided by the Virtual Writing Tutor on IELTS Writing Task 1 Essays: An Exploratory Study

Mashael Salem Alsalem
Department of English Language, College of Language Sciences,
King Saud University
Riyadh, Saudi Arabia
Email: malsalem@ksu.edu.sa

Received:04/16/2026 Accepted: 07/01/2026 Published: 07/15/2026

Abstract
This study explores the nature of automated scoring and feedback provided by an artificial intelligence-powered tool, the Virtual Writing Tutor. It also assessed the quality of the feedback specifically for non-native writers of IELTS Writing Task 1 essays. The study therefore contributes to the general research literature devoted to evaluating writing aids, which as yet has paid scant attention to this particular tool, and relatively little to the actual quality of feedback on writing for specific tasks. Eight students participated in this convergent mixed-method study, conducted during a 30-hour IELTS preparation program. A quantitative student-based perspective came from a 15 item, 5-point Likert scale response, questionnaire administered to students concerning their subjective evaluation of the Tutor's feedback. A qualitative expert perspective came from researcher review of the automated scoring and feedback provided on the participants' IELTS Writing Task 1 practice essays. The findings showed that the Tutor provided feedback relevant to all four of the IELTS writing assessment criteria, albeit the feedback was not presented straightforwardly in the IELTS categories. The software often used indicators, rather than more direct measures, and so could be regarded as overlooking important details that a human evaluator would notice. Additionally, the Tutor provided wrong information in some specific lexical and grammatical correction feedback. Students however judged that the Tutor was easy to use and (surprisingly) significantly useful in helping them improve grammatical accuracy, but they did not find it definitely effective in terms of improving spelling, content, vocabulary, and variety of sentence structures.

Keywords: Artificial Intelligence, error correction, evaluation, feedback, IELTS, scoring criteria, Virtual Writing Tutor, writing

Cite as: Alsalem, M. S. (2026). Evaluating the Automated Scoring and Feedback provided by the Virtual Writing Tutor on IELTS Writing Task 1 Essays: An Exploratory Study. *Arab World English Journal (AWEJ) Special Issue on CALLVol.17.* (12).237-256.
https://dx.doi.org/10.24093/awej/call12.14

Introduction

The significant improvement in AI capabilities in recent years offers the promise of enhanced efficiency, scalability, and personalized support for EFL writers from use of Automated Scoring and Feedback Systems (ASFS) or Automated Writing Evaluation (AWE) (Polakova & Ivenz, 2024; Zakaria et al., 2025). However, the enthusiasm surrounding such developments should be tempered by rigorous evaluation. The present study contributes to that enterprise by focusing on a dedicated AI assisted writing aid called Virtual Writing Tutor (VWT). While much less researched than resources such as ChatGPT and Grammarly, it enjoys considerable currency around the world e.g. in Canada (where it was developed), Indonesia, and the Arabian peninsula (where the present study was conducted) (Al Badi et al., 2020).

Recent studies suggest that the objective effectiveness of AWEs, and user subjective perceptions of their usability, can differ significantly. They may further vary depending on factors such as the specific tool deployed, the target population using it, and the criteria used for evaluation (Almegren et al., 2024; Escalante et al., 2023). All these variables cannot be pursued in a single study. The present one chooses both an objective and a subjective perspective, VWT as the tool, and Saudi Arabian university level writers as the target population.

Prior evaluative research has highlighted that AWEs generally can enhance aspects of feedback on writing such as identifying grammatical errors and providing general suggestions (Polakova & Ivenz, 2024). However, concerns have been raised about AWE's ability to provide nuanced, context-aware guidance (Almegren et al., 2024), accuracy and reliability (Escalante et al., 2023), and a beneficial impact on students' emotions and learning (Chan et al., 2025). To address this gap in evaluative information about VWT, the present study focused on two key aspects affecting AWE success. Expert researcher judgment addressed objective accuracy of the scoring and feedback provided by VWT, and learner perceptions of its ease of use and effectiveness measured users' subjective view of its quality. I draw on Carroll's (1963) model of school learning, UTAUT software adoption theory (Venkatesh et al., 2016), and the DeLone and McLean model of information system success (Petter et al., 2008). The essence of this for the present study is shown in Figure 1. AWE benefits for writing depend both on the objective quality of a resource such as VWT and also on learners' perceptions of that quality. Specifically, student use of VWT is assumed to depend on students' perception of its quality (which may partly reflect its actual quality); student writing improvement is assumed to depend both on VWT use and on the objective quality of the information provided by VWT. The slight literature on VWT (covered below) shows its improvement effects may be limited. Thus, to evaluate and better understand VWT we need to explore both the VWT factors addressed in this paper: its perceived quality and objective quality.

The VWT is specifically designed to assist individuals in improving their English writing skills like those that the International English Language Testing System (IELTS) writing proficiency test requires. IELTS is widely used, especially in relation to study at tertiary level. Although the literature has explored the potential of AWE in preparing for various writing assessments, research has not examined VWT application in the context of preparing test-takers for standardized English language assessments, such as IELTS, nor has it focused specifically on the IELTS Writing Task 1 component. This study is therefore also innovative in assessing VWT specifically in relation to IELTS writing by tapping into learners' attitudes as well as expert opinion on the effectiveness of VWT.

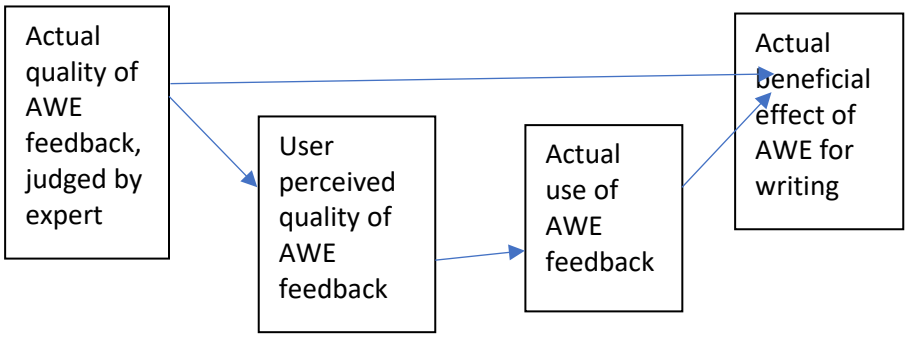


Figure 1. Assumed relations between key variables in AWE software innovation (simplified from Petter et al., 2008)

Research questions

In the context of tertiary level EFL writers completing IELTS Writing Task 1:

1. How easy to use and helpful do students perceive VWT to be?
2. What is the expert view of the nature and quality of VWT feedback to writers?

Review of Literature

Due to the relative paucity of studies of VWT, this review includes reference to key issues in studies of a wider range of digital writing aids. Broadly, Automated Writing Evaluation (AWE) systems provide immediate or post draft feedback on what a writer types, in the form of scores and/or comments. They potentially evaluate at all levels of language, and indeed content, not just mechanics, grammar, and vocabulary. They may identify not just errors but qualities such as diversity or interest (Zhang & Hyland, 2022). In this way, AWEs, like human feedback givers, contribute what is widely regarded as a key requisite for language learning (SLA), responses that draw the writer's attention to errors and other features of their output. Only with such feedback do learners 'notice', discover corrections, and so bridge the gap between their current L2 proficiency and their target level. Effective L2 feedback, whether explicit correction or guided prompts, is essential not just for draft improvement but also long-term learning (Schmidt, 2010). Feedback also connects with testing in that it constitutes a type of formative assessment. That is feedback which is typically not just a summary final score representing how much has been achieved. Rather, it is assessment seen as a forward looking contribution to an ongoing learning process. For L2 writing, formative feedback can be delivered in a way that has wider benefits, reducing writing anxiety, and promoting self-regulation.

Clearly much of the above is traditionally delivered by human (teacher, peer) feedback, though probably not so exhaustively as by AWE, and certainly not instantly at the moment of writing. Importantly the above benefits arise in both cases on the often unstated assumption that the feedback is in fact correct.

Both before and after the recent blossoming of AI, researchers have done research that essentially compares AWEs with human feedback givers. Much of it compares the two by their effects on student writing. AWE for example can significantly improve writing proficiency by reducing grammatical and lexical errors. This has been shown both for improvement over drafts of a single essay (Polakova & Ivenz, 2024) and for improvement over time on different essays

(Jamshed et al., 2024). Wei et al. (2023) assessed the effects of using the AI-driven platform, Grammarly, on student achievement in the four areas targeted by the IELTS scoring system (task achievement, coherence and cohesion, lexicon, and grammatical accuracy). The control variables for this study included writing self-efficacy and proficiency in global English. The results showed that in all aspects of writing, the experimental group consistently demonstrated better performance than the control group (which received some human feedback).

Research has also been done on learners' perception of feedback. Generally they express positive attitudes toward AI-assisted feedback, as they associate it with noticeable improvements in their writing skills (Jamshed et al., 2024; Kaharuddin et al., 2024). This information is valuable as it gives an idea of whether the students would obtain and use AWE feedback when left to themselves and not participating in a research study where they are usually required to use it.

Affective impact has also been noted. Mohammed & Khalid (2025) conducted a mixed-method study to examine the effect of Al-generated feedback and found that it improved motivation, foreign language peace of mind, emotional intelligence, and quality of writing. Chan et al. (2025) explored the way ChatGPT in a Hong Kong university found that students receiving Al-generated feedback not only demonstrated statistically significant improvements in quality of writing but also higher motivation and engagement.

The findings for AWE are not always positive however. Escalante et al. (2023) explored the impact of AI-generated feedback on writing in a longitudinal, quasi-experimental study, However, the study found no significant linguistic progress in these groups, and preferences for AI-generated and human feedback were mixed. Chan et al. (2025) found that, although ChatGPT improved quality of writing and learners' engagement, study participants demonstrated mixed feelings about its effectiveness. In addition objections have been raised concerning the risk of over-reliance on AWE, concerning its reliability and fairness, and that it may hinder the development of learners' personalized critical thinking and writing skills (Behzad et al., 2024).

Furthermore, many experimental studies of the effects of AWEs over time such as that of Wei et al. (2023) suffer from objections such as that the content of the feedback, not just the delivery method, was different between the experimental (AWE) group and control group. Hence it is not clear if any effects were due to the content or the delivery vehicle. Furthermore the AWE was often something new so could have benefited simply from novelty effect. In addition the actual nature and quality of the feedback (a focus of the present study) often has not received attention, only later writing products that are assumed to have been influenced by the prior feedback. Finally, in academic settings writing is usually done with the requirement to satisfy some detailed prompt or scoring rubric (e.g. that of the English Dept, or of IELTS etc). Researchers however rarely seem to take feedback quality and specific writing goals into account: the present study remedies that.

One study that did compare the actual AWE-generated feedback with that provided by human instructors or other human sources was Almegren et al. (2024). This compared AI-generated feedback (ChatGPT-3, Gemini, and Bing) with human feedback across multiple essay evaluation categories. The findings suggested that human raters appreciated qualitative features of the writing better than AIs. ChatGPT performed better in vocabulary and content analysis than Bing and Gemini. The researchers highlighted the need to refine AI use and suggested that, although AI had potential, it required hybrid application with both humans and Al.

Indeed hybrid frameworks are now quite widely proposed that combine the instant, multi-level suggestions of AWEs with the empathy, context, and nuanced oversight of human instructors

(Altamimi, 20i4). They encourage students to actively evaluate AI suggestions rather than blindly accept them. Escalante et al. (2023) also suggested blending both AI-generated and human feedback.

Only a few studies have evaluated VWT, mostly in minor journals. They reflect similar findings to those above. Nasution et al. (2023) evaluated the perceptions of students when using VWT feedback on writing at a university in Indonesia. While positive reactions highlighted the nature of the feedback and the way VWT encouraged independent learning, negative reactions concerned features not central to the VWT feedback itself, such as the lack of options in essay topic, and the paid nature of some VWT features. Viantika & Dangin (2024) found that, although VWT improved high school students' EFL writing in Indonesia, not all interview comments about it were positive. Zein et al. (2025) tested VWT among university students in Indonesia and found that it helped EFL students to "generate, structure, and refine their ideas with more autonomy. Students also showed improved awareness of language mechanics and vocabulary selection." Interestingly, they also found that it complemented teacher feedback, implying that the VWT was not self-sufficient.

Al Badi et al. (2020) in Oman align more closely with the context of the present study (in Saudi Arabia). They found that VWT improved spelling, punctuation, grammar, and vocabulary of student writing at an applied science college. However, the findings of this study were not comparable with those of the present study, which examines the actual nature of the feedback and targets more proficient students writing for IELTS.

Overall, as in the wider literature, there is a dearth of VWT studies which evaluate the actual quality of the scoring and feedback for writers at higher proficiency levels, specifically in relation to high-stake examinations such as IELTS with its distinctive scoring categories. This is despite the insistence of educational experts such as Hattie and Timperley (2007) that the actual nature and quality of the feedback is supremely important. Furthermore the student perception of the feedback (not just the reality), although already tapped in nearby Oman, also needed to be ascertained, as emphasized in the wider field of technology acceptance (e.g. Venkatesh et al, 2016): clearly it may affect use. Hence the nature, quality and student view of feedback are the foci of the present study, with the added feature that a particular writing task is focused, not writing in a general undefined sense as in many studies. This makes the present study a unique contribution. The nearest I can find to the present study is Zhao et al. (2026).

Method

This convergent mixed-method study used both qualitative and quantitative data to gain insights into the nature of writing feedback provided by VWT, and student perceptions of it.

Participants

Participants were eight faculty members, from a variety of disciplines at a university in Saudi Arabia, enrolled in an IELTS preparation course (Table 1). All participants were Saudi nationals with English proficiency IELTS 6 or more, so CEFR B2 or higher. The size of the sample, like its membership, was determined by convenience.

Table 1. *Participant demographics*

#	Age	Degree	Major	Gender
1.	27	BS	Medical Engineering	M
2.	27	BA	International Finance	F

3.	32	MA	Business Administration	F
4.	36	MS	Biology	F
5.	26	BA	Art	F
6.	31	BS	Respiratory Therapy	M
7.	31	BA	Special Education	F
8.	38	MS	Chemistry	F

Research Instruments

Two methods were used to gather data for evaluating VWT (Calhoun et al., 2025). Firstly, a closed quantitative questionnaire was used to obtain participants' subjective perceptions of the quality of VWT feedback. A 15-item questionnaire was developed to elicit the degree of agreement or disagreement with statements referring favorably to aspects of the use of VWT on a 5-point Likert scale: 5 strongly agree, 4 agree, 3 neutral, 2 disagree and 1 strongly disagree. The following domains were covered (with number of items for each): time benefits (3); technical ease (2); overall feedback clarity, accuracy, and completeness (4); feedback value in relation to aspects of writing scored by IELTS (6: content, text structure, coherence, grammar accuracy, vocabulary and grammar variety, and spelling). The items are listed and classified in Table 2 in Appendix. This questionnaire differed from that used by Al-Badi et al. (2020), as their study attempted to measure the motivational effect of VWT (aspects such as liking, perceived difficulty, interest etc.), whereas this study focused on how participants perceived the quality of VWT feedback for a specific writing task.

Secondly, expert review (Morrison et al., 2019), in the form of qualitative document analysis conducted by the researcher, was used to analyze the draft texts of tasks completed by the participants and the feedback and scores provided by VWT on those drafts. This ascertained the way the latter was associated with the former, and to the aspects indicated as crucial by their mention in the IELTS criteria.

Context and Procedure

All study participants were enrolled in a 30 hour IELTS academic preparation course (7.5 hours on writing).

The quality of writing in the IELTS examination is measured by two tasks, of which the relatively easier Task 1 is the focus of the present study. Task 1 requires examinees to understand the information provided via a table, graph, diagram, or image(s) and write a summary of that, in at least 150 words, making contrasts where relevant. Examinees are advised to spend 20 min on Task 1 (out of 60 min overall allocated to writing), and the score obtained is weighted half as much as Task 2. Each task is assessed on four criteria, all of which are weighted equally, and together, they comprehensively measure all relevant aspects of the text written. Each criterion is scored on a scale of 0 to 9, and the overall Task 1 score is the average of the four criteria (often rounded to .5), called the IELTS Band score (IELTS, 2025). The four criteria and their components are as follows:

1. Task Achievement (TA)
   - selecting key features of the information provided
   - providing sufficient detail to illustrate these features
   - reporting the information, figures, and trends accurately

▪ comparing or contrasting the information by adequately highlighting the identifiable trends, principal changes, or differences in data and other inputs (rather than mechanical description reporting detail)
▪ presenting the response in an appropriate format

2. Coherence and Cohesion (CC)
▪ the coherence of the response provided via the logical organization of information and/or ideas, or logical progression of the argument
▪ the appropriate use of paragraphs for topic organization and presentation
▪ the logical sequencing of ideas and/or information within and across paragraphs
▪ the flexible use of reference and substitution (e.g., definite articles, pronouns, etc.)
▪ the appropriate use of discourse markers to clearly mark the stages in a response, e.g., [First of all | In conclusion], and to signal the relationship between ideas and/or information, e.g., [as a result | similarly]
3. Lexical Resource (LR)
▪ the range of general words used (e.g., the use of synonyms to avoid repetition)
▪ the adequacy and appropriacy of the vocabulary (e.g., topic-specific items, indicators of writer's attitude, etc.)
▪ the precision of word choice and expression
▪ the control and use of collocations, idiomatic expressions, and sophisticated phrasing
▪ the density and communicative effect of errors in spelling
▪ the density and communicative effect of errors in word formation
4. Grammatical Range and Accuracy (GRA)
▪ the range and appropriacy of structures used in a given response (e.g., simple, compound, and complex sentences)
▪ the accuracy of simple, compound, and complex sentences
▪ the density and communicative effect of grammatical errors
▪ the accurate and appropriate use of punctuation

Participants received three hours of traditional instruction, where the trainer explained the nature of each writing task and asked participants to collectively compose one on the board. Thereafter, the participants were handed another task, to complete the task by themselves. The participants then practiced further on materials relevant to Task 1 writing before VWT was introduced and explained to them. They were asked to create their VWT accounts and after a period of familiarization with it, to attempt a new IELTS Task 1 by typing directly into the onscreen box provided by the VWT, which could provide score and feedback for each task when requested. Each participant had 20 min to complete their task, and no resources other than the VWT were allowed to be used. After participants had completed the task, they were asked to obtain the VWT score and feedback and send those to the trainer via email. Upon completion of this process, the participants were asked to complete the questionnaire described above.

The VWT Tool

The VWT (https://virtualwritingtutor.com/) is a free, online grammar checker and writing feedback engine designed to help language learners improve their writing skills (the paid version was not used in this study). Although it caters to a general audience, it has been claimed that features of this tool are specifically tailored to assist students preparing for standardized English language tests, namely IELTS (Virtual Writing Tutor, 2025). While the tool is claimed to be AI

driven, it also uses some traditional automated measures of written text, such as average sentence length. Alongside correcting and enhancing the text, the tool also provides scores and comments on the changes made. The ability of VWT to improve grammar, vocabulary, spelling, and paragraph structure is highlighted on its webpage (esp. IELTS criteria, 2-4). However, its ability to improve the content with respect to the writing prompt is not emphasized (IELTS criterion, 1).

Data Analysis

Questionnaire data was collated and analyzed numerically in Excel. The binomial test was used to determine the items that exhibited responses differing significantly from a 50-50 agree–disagree split of opinion (omitting those who responded 'neutral'). The expert review data consisted of pairs of texts to analyze: each student Task 1 draft text had a corresponding VWT text with scores and feedback. These pairs were submitted for qualitative analysis. The researcher read over the pairs repeatedly to ascertain which IELTS aspects were covered and how effective the feedback was for the participants. Essentially this was a comparison of what VWT provided with what the students needed (=what the IELTS criteria valued): such a comparison lies at the heart of many systems for evaluating educational resources (Atigh & Khabbazi, 2021).

Results

Questionnaire to participants

The quantitative questionnaire data (Table 2 in Appendix) showed that the majority of the statements (80%) elicited mean agreement with positive features of VWT that was above the neutral midpoint judgment (3 on the agreement scale). Of those, more than half (58.3%) were significantly above the neutral rating, as determined by the 50% binomial test ($p < .05$). Although three statements (20%) were rated below neutral 3, they did not reach the level of significance. Over all 15 items, the average endorsement of good quality is 3.7. That is above the neutral midpoint (3) but falls short of 'agree' (4).

After examining the substance of the statements, it was found that the seven statements which were significantly endorsed by the participants predominantly involved features that were not textual aspects of writing specifically. Rather, they underscored VWT benefits for time/speed, its technical simplicity, and clarity of the feedback. Only one of those significantly positive statements focused on improving a specific IELTS relevant textual aspect in writing - grammatical accuracy.

In the range of scores that were descriptively positive (average above 3, but not significantly so) fell approval of VWT's help with completeness, accuracy and effectiveness.

Also in that range fell two more of the IELTS relevant items: text structure and coherence (both discourse features wider than the sentence).

Then with averages showing endorsement below the midpoint (but not significantly so) came VWT's help with IELTS relevant items: vocabulary (lexis) and sentence variety (i.e., grammar variety), content (i.e., task achievement), and spelling.

Expert review of the feedback

The scores and feedback for all texts submitted to VWT appeared under the same set of headings. and subheadings, in a prearranged order (Table 3), with all categories listed regardless of whether there was any feedback given. The correspondence with the IELTS four categories

(representing student needs) was however complex and unstated. The account below closely follows the VWT categories.

Table 2. *VWT reporting categories in relation to IELTS criteria*

VWT Category of feedback/score provided	IELTS assessment categories that appear to be addressed most closely
Band scores	Band scores
Writing fluency	Requirement of at least 150 words in 20 minutes (7.5 wpm)
Writing style: Average sentence length	GRA-complexity
Writing style: Sentence length variance	GRA-range
Writing style: Average sentiment strength	LR-style, TA
Writing style: Cliches detected	LR-style, TA
Writing style: Number of exclamation marks	GRA-style, TA
Organization: Paragraphs	CC
Organization: Cohesion	CC-range
Topic development: Focus	TA, LR-range
Topic development: Depth	TA, LR-range
Topic development: Control	TA, LR-range
Language accuracy/Number of errors	LR-accuracy, GRA-accuracy (incl. spelling and punctuation)
Grammar check feedback	

Band Scores: VWT approach and Comparison with IELTS

VWT used 13 feedback categories, of which 12 were scored on a 0-9 scale and an overall Band score derived from that. This contrasts with IELTS where the band score is the average of component scores for its four criteria: TA, CC, LR, and GRA. The outcome band scores provided by VWT were judged as plausible by the researcher based on experience (albeit she was not a trained IELTS examiner), but with a tendency to err on the high side for reasons that will appear below (some weaknesses not detected).

User suitability

The disadvantage of VWT for students is that most IELTS preparation teaching and textbooks reflect the 4 categories followed by IELTS. Hence, students approaching IELTS through that route, such as ours, were familiar with the IELTS four categories for writing tasks and might find it easier to process component scores and feedback presented under the four IELTS criteria, rather than the 13 categories of VWT. Furthermore, the tendency to score on the high side may encourage false confidence. Traditional teachers are often mark examination practice tasks on the low side, as they do not want to encourage students to develop the impression that they do not need to improve before an actual examination.

Fluency: VWT approach and Comparison with IELTS

The Fluency category in VWT reports the total number of words written and provides a separate comment and score: e.g. 6 points out of 9 for 154 words and 9 points for 191 words. Although IELTS does not separately score the number of words written, it assesses them words written indirectly via its four assessment categories. IELTS, in effect, recognizes that writing more may have both good and bad effects. More words can facilitate TA (more complete coverage, comparisons etc.), CC (if paragraphs are logically connected), LR (chance to use more vocabulary), and GRA (opportunity to use different sentence structures). However, longer texts also increase chances of repetition (hurting CC), off-topic content (hurting TA), and grammar or

vocabulary mistakes (hurting LR and GRA). Furthermore, writing more for Task 1 may damage students' performance on Task 2, which carries more marks; therefore, time management is necessary to ensure that no more than the 20 min allocated to Task 1 is used.

User suitability

As discussed above, higher scores based purely on the length of writing may not be beneficial for students. Furthermore, it is unclear whether VWT takes time into account. Although it records time spent on writing, it reports that quite separately from the way it reports Fluency. However, fluency is usually defined in terms of the numbers of words written per minute. The VWT comments appear to refer only to the numbers of words. For instance, a participant who wrote 154 words received the following comment: "You have just enough words. Try to write a few more." Given that the timer on the page recorded 10:11, the following comment might be more pertinent: "You have just enough words, but you have 9 minutes available to write more." Furthermore, given that the Average sentence length report lower down in the comments suggests some longer sentences would benefit the score, the following comment could be added: "Why not try lengthening some of the sentences to improve your grammar score (see below)?" Therefore, there are signs that VWT generates Fluency responses in an isolated, self-contained manner, in which only length is focused upon, and does not develop an association among length, time, and other feedback, which may benefit students from writing more.

Content: Topic (3 VWT categories) compared with Task Achievement (IELTS TA) and Vocabulary Range (part of IELTS LR)

The three VWT topic categories (Table 3) are designed to deal with this and refer to the "chart" (graph, table, or image), which acts as a guide to the content that is expected in Task 1. The three VWT components are "Topic development: Focus," which concerns "the concepts in the chart"; "Topic development: Depth," which concerns "the vocabulary a reader is likely to expect in the analysis of a chart like this"; and "Topic development: Control," which concerns "the information contained in the chart." Responses in each of these three categories consist of a list of words and phrases that VWT predicts should occur in writing for the specific task and a count of how many of those a writer has actually used, with a score. A higher score is awarded for using higher numbers of words and phrases. For instance, using 23 of 84 words and phrases or 32 of 57 words and phrases yielded a score of 9. Any vocabulary prediction like that is only feasible for a very constrained topic, and the IELTS Writing Task 1 is indeed highly constrained by whatever information is displayed in the graph or table given for each task (e.g., a diagram of the steps that bees follow to make honey). By contrast it would not be easy to predict the vocabulary a competent student will use to write an essay on a prompt, such as "My favourite hobby."

IELTS approaches content requirements in quite a different way. Rather than specifying vocabulary expectations that should be met, it talks about "selecting key features of the information, providing sufficient detail to illustrate these features, reporting the information, figures and trends accurately, comparing or contrasting the information by adequately highlighting the identifiable trends, principal changes or differences in the data and other inputs (rather than mechanical description reporting detail), presenting the response in an appropriate format" (IELTS, 2025). In short, IELTS conceptualizes the content required in terms of speech acts and information packaging, which are at the heart of content, rather than reducing it to the variety of

vocabulary that may be used to express that (this concept of content falls under IELTS LR, as a range).

User suitability

The differences in VWT and IELTS approaches to content suggest that meeting the VWT feedback has little connection with whether the IELTS TA requirements are met. Therefore, using VWT does not necessarily help a writer to improve what truly constitutes TA and may produce a falsely optimistic score. The issue arises because, although students may use a good range of suitable words (to obtain high score in VWT), the information may not be accurate. For instance, although the phrases "adults improved more than children" and "children improved more than adults" use the same vocabulary, only one is correct, given the information supplied in a graph.

Furthermore, there are problems even just considering the VWT as providing vocabulary range feedback (IELTS-LR). The VWT vocabulary lists are long and burden the user with information that they cannot readily use. In a task where the stimulus was a bar chart about internet use over three years by different age groups in Taiwan, the numbers of items listed as relevant were: Focus 50, Depth 55, Control 63. Of these respectively one participant used 15, 32, 27 and was awarded a score of 9 for each. However, the lists of items do not separate which words the user did not use from those they did, so it is hard for the student to rapidly benefit from the lists. Furthermore the lists overlap considerably, making the information in them unnecessarily repetitive to read and process. In this instance the overlap of target words was: Focus and Depth 34; Depth and Control 40; Focus and Control 38. Indeed 30 items appeared in all three lists (examples: "indicate, display, provide, Internet, Taiwan, young, adults, old"). Again there could be benefit if those that the participant had not used were highlighted. In addition, an underlying issue is surely what real difference there is between what VWT calls Focus, Depth and Control and, if there is a difference, how it benefits the writer to receive feedback for those separately.

Organization (VWT) and Coherence (IELTS CC)

The Organization category of VWT, with two subcategories, corresponds to IELTS CC. VWT specifies the number of paragraphs that it thinks the text should contain and a list of "transition words" that it thinks should be used in the text. For example, in a task about popularity of different news media, VWT expects that text should contain 3 to 4 paragraphs, and it provides 12 transition items, including "whereas, although, thereafter, afterward, however, overall, on average, according to the information". VWT scores the text separately based on its closeness to VWT in number of paragraphs and number of listed transition words.

Although IELTS pays attention to both these aspects, it goes further by asking for logical organization and sequencing of information and/or ideas, flexible use of reference and substitution (e.g., pronouns), and the appropriate use of discourse markers. In other words, IELTS is concerned with coherence as connected meaning, not just as occurrence of the linguistic forms that mark the coherence.

User suitability

As IELTS is concerned primarily with coherence of meaning, the VWT approach, which does not consider whether "transitions" are used appropriately, falls short of meeting the needs of students, and may lead to overestimation of the IELTS score. For example, VWT may not detect and consider anomalous an example such as "Although Australia and Canada show similar

percentages accessing the news via TV, their use of online channels to access the news is almost the same". The VWT approach seems to be that if a student uses any of these "transitions", it is very likely they use it correctly so accuracy need not be checked. Similarly if they use an appropriate number of paragraphs, there is no need to check if the contents of those paragraphs is separate but connected as it should be.

Tone: Part of Style (VWT) and part of Vocabulary and Grammar (IELTS LR and GRA)

VWT has three subcategories which explicitly deal with Style from the perspective of tone. First, it reports Average sentiment strength, presumably by counting specific words or expressions in the text which appear in stored lists of words that have been classified as having positive or negative emotional or attitudinal connotations (e.g., calling a finding "magnificent" or "poor"). These convey a non-neutral meaning, which is not suitable for academic writing (covered by IELTS LR and TA). VWT calculates a sentiment score or a text, where lower values indicate better performance. However, VWT does not specify which words in the student text are regarded as excessively loaded. Therefore, users may fail to replace those and may not learn to avoid them in future. Furthermore, despite its overt attention to this aspect of Style, VWT surprisingly responded to a student text about food consumption which contained emotionally marked words, such as "dramatic, whopping, huge, glorious", with the comment: "It is important to maintain a neutral tone in academic writing and simply report the information without giving a strong opinion. You have done that. Good job."

Second, VWT treats exclamation marks as a separate style category, not just punctuation (IELTS GRA). They add emotional significance to the content, so are not suited to academic writing (IELTS TA). Therefore, using lower numbers of exclamation marks in the Task 1 essays is awarded higher scores by VWT. However, although easy to identify, this feature is quite rare and was absent from this study's data.

Third, VWT targets Cliches. However, this category is hard to define and apply consistently and arguably seems less important to target than informality for academic writing. Informal usage such as "online news, which makes up the majority.." or "The lowest band weren't consistent" are not picked out.

The three VWT style or tone categories have no exact counterparts in the IELTS approach, where they are not emphasized. For instance, in IELTS, "indicators of writer's attitude" are one of the list of features of LR and "appropriacy of structures" with "appropriate use of punctuation" fall within GRA. Thus, compared with IELTS, VWT gives more importance to style or tone. However, while VWT focuses on specific subtypes only, words, such as "appropriate" used by IELTS suggests that all style features need attention.

User suitability

VWT emphasizes its style features even in those cases where texts show minimal errors. Owing to this, the attention of a writer can be drawn away from factual matters of content that IELTS values (e.g., identifying main points, accuracy of what has been stated, and making comparisons). A typical VWT feedback is as follows:

> "Average sentiment strength: 0.2. Score: 9. Feedback: It is important to maintain a neutral tone in academic writing and simply report the information without giving a strong opinion. You have done that. Good job.

Number of clichés: 0 Clichés detected. Penalty: 0. Feedback: Keep your writing fresh by avoiding clichés.

Number of exclamation marks: 0. Penalty: 0. Feedback: Never use exclamation marks in academic writing."

Overall, although the focus of VWT on these stylistic features is not incorrect, it has several weaknesses for users. As items with unwanted marked sentiments are not indicated to the user, they cannot readily be improved. In addition, VWT's choice of clichés may be debatable, and it does not target several stylistic features (e.g., formality) of academic writing.

Grammatical Range: Part of Style (VWT) and of GRA (IELTS)

VWT provides information in the following manner:

"Writing Style Average sentence length: 20.22 Score: 9 Feedback: An appropriate average sentence length for this task is 25 words. Your sentences are the right length.

Sentence length variance: 4.92"

The average length measure is often used by scholars as an indicator of grammatical complexity, as longer sentences are usually more complex grammatically. The "variance" is an indicator of the degree of variety in grammatical structures that students used in writing (technically the figure provided is the standard deviation and not the statistical variance).

IELTS deals with this within GRA. However, unlike VWT, which focuses on sentence length, IELTS refers to simple, compound and complex sentences. The IELTS approach prefers to focus directly on the subject of interest, rather than an inexact indicator of that.

User suitability

The VWT information is useful as a rough guide. However, as VWT does not pay direct attention to different syntactic structures, it cannot fully deal with IELTS requirements.

Accuracy of Vocabulary, Grammar, Spelling, Punctuation: Language Accuracy and Grammar Check (VWT) compared with LR and GRA (IELTS)

VWT evaluates together the accuracy of a range of language forms that traditionally have often been the main focus of feedback: grammar, lexis, and mechanics (spelling and punctuation). Errors are counted to arrive at an overall score, and the feedback is provided in a three-part error correction format, as shown in the following example:

You wrote: ...ount of fast food consumption in the UK between 1970 to 1990. Throughout these 20 years, the graph s...

Feedback: Use "between" + "and" not "to." Try this: "between 1970 and 1990."

Suggestion: between 1970 and 1990.

However, the researcher identified several weaknesses in the error correction. First, all errors were not addressed. For instance, no feedback was offered on: "The park included a children playground and around pound in the middle of the part surrounded by many trees" (should be "… a children's playground and a round pond in the middle of the park…"). Of course, general wisdom on error correction suggests that experts may deliberately choose not to correct all errors in a text written by students because either they do not want to discourage the students or the current lesson does not focus on that error. However, when preparing quite proficient students for an examination, as in the present case, covering all errors is a defensible policy. Indeed VWT gives the impression that it attempts to capture all errors; otherwise, awarding "0 errors" to some texts will be without

meaning. One text declared by VWT as having no errors however begins as follows: “Countries have shown differences in the way they accessing news. The chats display two examples of how Canada and Australia accessing the news….” (possible corrections: simple present tense needed, or perhaps “are accessing”; “chats” actually refers to the “charts” provided).

Second, in contrast with not identifying real errors is identifying nonexistent errors, which also occurs, and sometimes the “correction” introduces an error. Example:

> You wrote: ... age groups exhibited a marked increase under 15 and over ... in internet adoption, with users
> Feedback: In English, since there is only one internet that everyone knows about we normally use a definite article, and say, “the internet.”
> Suggestion: the internet.

The following example shows that VWT has misinterpreted the sentence, as it considers “access” to be a verb, whereas, in this context, “reveals” is the verb. Although the sentence had no error, the “correction” by VWT produced one.

> You wrote: The comparison of news access in Canada and Australia reveals distinct preferences among the populations...
> Feedback: Your subject (The comparison) is singular, but you have used a verb conjugated for a plural subject. Revise “The comparison accesses.”
> Suggestion: The comparison accesses

Another possibility is that a genuine error is missed and a nonexistent error is identified together. In the following example, VWT appears to admit being unsure of the error and suggests consulting with a secondary resource (the actual error is that “it” should be “in”).

> You wrote: ... The graph shows the ages of internet users it Taiwan spanning over the years 1998, 1999 and ...
> Feedback: A sentence without a verb is called a fragment. Please, verify.

Furthermore, some errors that are correctly identified as errors are nevertheless misunderstood, and the correction suggested is also an error. Here, although an error is correctly identified by VWT in “The foundation of Sheman part”, the correction suggested is unacceptable. The real errors overlooked by VWT are: the prompt shows that the task is about “Sherman Park”; “foundation” does not collocate well with park, “creation” would be better; and “innovative” might be better than “innovated”.

> You wrote: ...d the town. The foundation of Sheman part marked the innovated development of the...
> Feedback: Your subject (The foundation) is singular, but you have used a verb conjugated for a plural subject. Revise “The foundation parts.”
> Suggestion: The foundation parts

In the following example, which shows a text about bees making honey, although the right correction (widely) has been provided in the feedback, it has been weakened by the fact that other meaningless options are provided. Furthermore, the feedback does not state that the phrase “in the form known widely” needs to be changed to “in the widely known form”.

> You wrote: ... needs to be produced in the form known widly. The process includes seven steps that ...
> Feedback: Possible spelling mistake found
> Suggestion: widely, wildly, Willy, wily, idly, willy, w idly

Clearly the IELTS criteria of language accuracy (in LR and GRA) are not met by the above.

User suitability

The inaccuracy, incompleteness, and ambiguity of multiple suggestions in areas that learners often pay most attention to will not meet their needs either for immediate correction or long term learning. Furthermore, the VWT way of citing portions of student text bounded by parts of words, rather than whole words or complete clauses, does not help users to rapidly understand feedback suggestions. Finally, the habit of in effect offering the suggested correction twice, which happens frequently, is unnecessarily repetitive.

Discussion

How easy to use and helpful do students perceive VWT to be?

VWT was rated higher on technical aspects, such as user friendliness, interface, and navigation, than on the core deliverables, such as improving quality of written text. In fact it is not uncommon for this to happen in surveys of attitudes to new software/IT in the workplace. In the Philippines, this was demonstrated in a similar study, which evaluated the use of Grammarly in EFL writing (Ventayen & Orlanda-Ventayen, 2018). The tendency to use software that is superior in terms of usability but inferior in terms of substantive help could however have a negative impact on students (Bringman-Rodenbarger & Hortsch, 2020).

In this study, IELTS Task 1 writing was the target and VWT feedback specifically relevant to that was mostly rated middling or low. Below the midpoint fell spelling, content, and vocabulary and grammar variety. Feedback on coherence and text structure were rated as positive but without significance, with only grammatical accuracy receiving a positive rating with significance (which accords for example with positive findings for grammar in Zakaria et al.'s 2025 survey, but not with our expert review of VWT). We cannot be certain about the reason for the low or intermediate ratings. The participants either were aware that the feedback was not always adequate or possibly felt they were already good in these areas, and VWT had little scope to help them improve further. For minimum B2 level speakers of English that could explain the finding for spelling but not so easily the others. To be fair, feedback on content was not highlighted by VWT itself as a strength but it equates effectively with criterion 1 (TA) in the IELTS scoring so its low rating is unfortunate.

What is the expert view of the nature and quality of VWT feedback to writers?

Expert review reveals a number of weaknesses, where the target student needs would not be met, if students were to read, and attempt to use optimally, all that is provided. On the one hand there is feedback (both scoring and correction) in all four IELTS main categories and many subcategories. On the other hand, however, there is: lack of transparency in how VWT categories map onto IELTS categories; attention primarily to overtly marked aspects of the text (e.g. vocabulary and length) rather than the essence of what is said; distraction provided by truncated quotation and repetition of minor points; inaccuracy of some vocab/grammar/mechanics corrections.

One reason, perhaps, is that VWT seems to make modest use of AI. Much of the feedback provided by VWT appears based on AWE driven by traditional computing, rather than AI. Compared with VWT, a fully AI-powered resource such as ChatGPT would have better understood and applied the IELTS criteria in most areas. However, VWT often appears to be relying on simple connections between superficial indicators and features of interest (e.g., vocabulary for topic and

style, punctuation for style, sentence length for grammar complexity, etc.). This dependence however may have changed between when the data was gathered and the present.

One corollary of the limited use of AI is that VWT fails to function at par in contexts where overt markers are subtle or distant, so non-AI pattern matching routines fail to parse or decode correctly. For instance, in the supposed error discussed above in "The comparison of news access in Canada and Australia reveals distinct preferences", VWT software correctly identified that "comparison" and not "news" was the subject but failed to spot that the verb that agreed with it was "reveal" and not the nearby "access" (which was. in fact, a noun and not a verb, in this example).

A further consequence of that, consistent with general literature on computer-aided writing assistance (e.g. Wei et al,. 2023), is that VWT would be expected be more successful in actually helping students with lower proficiency than ours (cf. Viantika & Dangin, 2024). That is because in lower proficiency writing, for example, simpler sentences are used where the subject and its verb come much closer together so VWT's approach can work more successfully. That could in turn also explain why Almegren at al. (2024) found that AI gave EFL students more feedback than humans and scored texts lower than humans. Our findings agree that VWT can give more feedback than typical humans, but finds the scores maybe a little higher due to errors missed.

Objective-subjective correspondence

Comparing the information from the review and the findings of the questionnaire, this study shows an incomplete match. However, it is not to be expected that student evaluation of a resource will perfectly reflect its actual quality. The negative response on spelling could have occurred due to VWT's performance on examples like "widely" described above. The negative response about content is consistent with VWT's limited ability to deal with content in the full sense. Surprisingly, VWT information on variety of vocabulary and grammar was not highly rated, despite the long lists of vocabulary provided by VWT (under Topic). Although the participants rated VWT's coherence-related feedback comparatively better, that was without significance, perhaps because VWT only addressed connectors and not pronominal connectivity or covert text structures.

Furthermore, given the number of mistakes in the VWT feedback, the moderately positive (though not significant) ratings that VWT received for overall completeness, accuracy, and effectiveness, and the significantly high rating for grammatical usefulness were surprising. Based on these ratings, it can be concluded that the participants, despite their proficiency, did not identify all instances of faulty or limited feedback which was found in the expert review. Alternatively, it could be assumed that they were being polite to the researcher, or evidencing "novelty effect."

Applying the assumed theoretical framework in Figure 1 (Petter et al., 2008; Venkatesh et al. 2016), the conclusion is that students left to their own devices outside of a research intervention might chose to use VWT primarily for grammar improvement, if at all. They would however risk processing a certain amount of actual disinformation and be better assisted by some of the other types of information offered that are limited, but not wrong.

Conclusion

In answering the research questions, it was found that VWT covers a range of distinct features of writing. Indeed, VWT feedback should help writers to improve in certain areas that are evaluated by IELTS examiners. However, VWT demonstrates significant limitations in its

potential ability to help students. Following the VWT suggestions, it is likely that writers' essays and sentences would become longer; more varied in vocabulary, grammatical structure, and connectors; more impersonal in style; and more accurate in some low level features. However, these benefits will help writers increase the IELTS score only up to a certain point. In reality, they will not necessarily improve in areas valued by IELTS, such as clarity of information handling; main-supporting idea separation; factual accuracy, logic, description of contrasts, trends; the coherent organization of information and/or ideas; the appropriate use of discourse markers; the precision of word choice and expression; the control and use of collocations; idiomatic expressions and sophisticated phrasing; and the accuracy of simple, compound, and complex sentences (IELTS, 2025). Furthermore, in several cases, VWT feedback may mislead the user, if it has not been verified by a human user.

This study is limited both in terms of the number of participants and range of evaluation methods used. In particular it did not measure actual benefit as improvements in essay drafts between before and after receiving VWT feedback. It assumed rather than tested the model in Figure 1. However, this study showed perhaps why existing studies of essay drafting with VWT did not report unequivocally positive results. More widely, studies are needed on AWE feedback among a similar group of writers, extending to cover together all four of the central variables (Figure 1). In addition, further studies are required to examine writer awareness of the possible limitations of AWE feedback.

Overall, the mismatch between student and expert views suggests that studies should not conclusively rely on student rather than expert views of quality as evidence for the feedback provided by the software. Furthermore, the 'hybrid' idea that the recent literature promotes about combining computer-based and human feedback must be addressed in more detail (Altamimi, 2014). The present study implies that the expert or teacher should not give feedback in parallel with VWT. Rather, the expert or teacher needs to intervene between VWT and the student to delete, correct, or supplement unsuitable VWT feedback, which contradicts one of the selling points of automated feedback, its immediacy.

Funding: This research is not funded.
Acknowledgments: Not applicable.
Conflicts of Interest: The authors declares no conflict of interest.
Authenticity: This manuscript is an original work
Artificial Intelligence Statement: AI and AI-assisted technologies were not used.

About the author

Dr. Mashael Alsalem is an assistant professor at the Faculty of Language Sciences, King Saud University. She has been teaching English for about 22 years and training on IELTS for four years. Her main research interests include AI in language teaching and learning, Discourse Analysis, and language teacher cognition. ORCID ID: https://orcid.org/0009-0004-7067-1085

References

Al Badi, A., Osman, M., & Al-Mekhlafi, A. (2020). The impact of Virtual Writing Tutor on writing skills and attitudes of Omani college students. *Journal of Education and Development, 4*(3), 101-116. https://doi.org/10.20849/jed.v4i3.828

Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., & Almegren, R. M. (2024). Evaluating the quality of AI feedback: A comparative study of AI and human essay grading. *Innovations in Education and Teaching International*, 1–16. https://doi.org/10.1080/14703297.2024.2437122

Altamimi, O. (2014). Dynamic Written Corrective Feedback revisited: A critique and an improved model. *Arab World English Journal, 5*(3), 180-189. https://awej.org/wp-content/uploads/2021/01/14-1-1.pdf

Atigh, R. & Khabbazi, S. (2021). The evaluation of ESP Textbooks based on students' perspectives. *International Journal of Linguistics, Literature and Translation, 4*(5), 56-62. DOI: 10.32996/ijllt

Behzad, S., Kashefi, O., & Somasundaran, S. (2024). Assessing online writing feedback resources: Generative AI vs. good samaritans. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1638–1644. Torino, Italia: ELRA and ICCL.

Bringman-Rodenbarger L., & Hortsch M. (2020). How students choose E-learning resources: The importance of ease, familiarity, and convenience. *FASEB Bioadvances, 6*(5), 286-295. doi: 10.1096/fba.2019-00094.

Calhoun, C., Sahay, S., & Wilson, M. (2025). *Instructional design evaluation*. LibreText Social Sciences. https://socialsci.libretexts.org/Bookshelves/Education_and_Professional_Development/Design_for_Learning

Chan, S., Lo, N. & Wong, A. (2025). Leveraging generative AI for enhancing university-level English writing: Comparative insights on automated feedback and student engagement. *Cogent Education, 12*(1), 2440182. DOI: 10.1080/2331186X.2024.2440182.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, 57.

Hattie, J. & Timperley, H. (2008). The Power of Feedback. *Review of Educational Research, 77*(1). https://doi.org/10.3102/003465430298487

IELTS (2025). *Understanding IELTS scoring*. https://ielts.org/organisations/ ielts-for-organisations/understanding-ielts-scoring

Jamshed, M., Ahmed, A.S.M.M., Sarfaraj, M., & Warda, W.U. (2024). The impact of ChatGPT on English language learners' writing skills: An assessment of AI feedback on Mobile. *International Journal of Interactive Mobile Technologies (iJIM), 18*(19), 18–36. DOI:10.3991/ijim.v18i19.50361

Kaharuddin, A., D., Mardiana., L. I., Arafah, B., & Suryadi, R. (2024). Defining the role of Artificial Intelligence in improving English writing skills among Indonesian students. *Journal of Language Teaching and Research, 15*(2), 568-678.

Mohammed, S., & Khalid, M. (2025). Under the world of AI-generated feedback on writing: Mirroring motivation, foreign language peace of mind, trait emotional intelligence, and writing development. *Language Testing in Asia. 15*. 10.1186/s40468-025-00343-2.

Morrison, G. R., Ross, S. J., Morrison, J. R., & Kalman, H. K. (2019). *Designing effective instruction*. Wiley.

Nasution, R., Debora, M., Saragih, B., Gultom. J., & Lubis, S. (2023). Virtual Writing Tutor (VWT) on students' writing skill improvement at writing class, *Asian TESOL Journal, 3*(1), 15-18. DOI:10.35307/asiantj.v3i1.42

Petter, S., DeLone, W. & McLean, E. (2008). Measuring information systems success: Models, dimensions, measures, and interrelationships. *European Journal of Information Systems 17*, 236–263 (2008). https://doi.org/10.1057/ejis.2008.15

Polakova, P., & Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2410101

Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker, *Proceedings of CLaSIC 2010, Singapore, December 2-4* (pp. 721-737). Singapore: National University of Singapore, Centre for Language Studies.

Venkatesh, V., Thong, J. & Xu, X. (2016). Unified Theory of Acceptance and Use of Technology: A synthesis and the road ahead. *Journal of the Association for Information Systems, 17* (5), 328-376.

Ventayen, R. J. M., & Orlanda-Ventayen, C. C. (2018). Graduate students' perspective on the usability of Grammarly® in one ASEAN state university. *Asian ESP Journal, 14*(7.2). http://www.scopus.com/inward/record.url?eid=2-s2.0 85061485593&partnerID=MN8TOARS

Viantika, S., & Dangin, D. (2024). Improving students' writing skill using Virtual Writing Tutor: Automatic corrective feedback. *EDULIA: English Education, Linguistic and Art Journal, 4*(2), 138–151. https://doi.org/10.31539/edulia.v4i2.8890

Virtual Writing Tutor (2025). https://virtualwritingtutor.com/

Wei P, Wang X, & Dong H. (2023), The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. *Frontiers in Psychology, 14*, 1249991. doi: 10.3389/fpsyg.2023.1249991.

Zakaria, N., Hashim, H., & Jamaludin, K. (2025). Mobile technology for enhanced ESL writing education: A systematic review of benefits and pedagogical innovations. *Arab World English Journal (AWEJ) Special Issue on CALL, 11*, 210-222. https://dx.doi.org/10.24093/awej/call11.13

Zein, T., Sinar, S., & Deliana D. (2025). Using Virtual Writing Tutor in teaching writing to EFL learners**.** *Jurnal Ilmiah Ilmu Terapan Universitas Jambi (JIITUJ), 9*(2). DOI: 10.22437/jiituj.v9i2.36052

Zhang, Z.V. & Hyland, K. (2022). Fostering student engagement with feedback: An integrated approach. *Assessing Writing, 51*, 100586. https://doi.org/10.1016/j.asw.2021.100586.

Zhao, H., Huang, J. & Whipple, P. (2026), Reliability of ChatGPT's scoring and actionability of its feedback in IELTS writing preparation classrooms: Teacher and student acceptance in China. *AERA Open*. June 2026. https://doi.org/10.1177/23328584261448

APPENDIX

Table 3. *User questionnaire results (* indicates p < .05 on the 50% binomial test applied to those that expressed an opinion other than neutral)*

			Agreement (n)					Mean
#	Statement	Domain	1	2	3	4	5	(1-5)
1	*The Virtual Writing Tutor was easy to navigate and use during my IELTS writing task 1	Technical ease				5	3	4.38
11	*The feedback provided by the Virtual Writing Tutor was clear and easy to understand.	Overall feedback clarity				5	3	4.38
3	*The platform's interface was user-friendly and did not require much technical skill.	Technical ease				6	2	4.25
4	*The Virtual Writing Tutor enabled me to complete writing practice sessions more quickly.	Time				7	1	4.13
7	*The AI tool provided useful feedback on my grammar mistakes.	Grammar accuracy				7	1	4.13
13	*Regular use of the Virtual Writing Tutor increased my confidence in completing IELTS Writing Task 1.	Time				7	1	4.13
2	*Using the AI tool helped me manage my time more effectively when practicing writing tasks.	Time				8		4.00
6	The Virtual Writing Tutor helped me improve the overall structure of my writing.	Text structure			3	4	1	3.75
5	I felt confident using the AI tool without needing external assistance.	Overall feedback complete-ness			4	3	1	3.63
14	The feedback from the AI was accurate and helpful in improving my writing skills.	Overall feedback accuracy			3	5		3.63
15	Overall, I am satisfied with the effectiveness of the Virtual Writing Tutor in preparing me for IELTS Writing Task 1.	Overall feedback effectiveness		1	1	6		3.63
9	The platform offered effective suggestions to make my sentences more coherent.	Coherence			4	4		3.50
10	I noticed a marked improvement in my use of vocabulary and sentence variety through the tool's feedback.	Vocab and grammar variety		3	4	1		2.75
12	The AI tool helped me develop more relevant and focused content in my writing.	Content		2	6			2.75
8	Using the Virtual Writing Tutor improved my spelling accuracy in writing tasks.	Spelling accuracy		5	2	1		2.50