Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

June 2022, Volume 26, Issue 2
pp. 78–105

ARTICLE

L2 learners' engagement with automated feedback: An eye-tracking study

Sha Liu, University of Bristol

Guoxing Yu, University of Bristol

Abstract

This study used eye-tracking, in combination with stimulated recalls and reflective journals, to investigate L2 learners' engagement with automated feedback and the impact of feedback explicitness and accuracy on their engagement. Twenty-four Chinese EFL learners revised their writing through Write & Improve with Cambridge, a new automated writing evaluation system that generates automated feedback with three different levels of explicitness. Data from multiple perspectives were collected and examined, including participants' eye movements, their stimulated recalls, and their responses/revisions to automated feedback on their multiple drafts. The results revealed that participants spent significantly more time and expended more cognitive effort in processing indirect than direct feedback. However, a lower percentage of indirect feedback was taken up, and the revisions participants made based on such feedback were less successful. These findings suggest feedback explicitness as a determining factor affecting learners' engagement with automated feedback and point to the need for timely, supplemental teacher or peer scaffolding in addition to automated feedback. The results also suggest that AWE tools need to be constantly updated to improve their feedback accuracy, as error-prone feedback may cause participants to make inaccurate amendments to their writing. In addition, teachers should help learners confirm the accuracy of AWE feedback.

Keywords: *Explicitness of Automated Feedback, Accuracy of Automated Feedback, L2 Learner Engagement, Eye-tracking*

Language(s) Learned in This Study: *English*

APA Citation: 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

Introduction

The past decade has seen rapid developments of more powerful and sophisticated Automated Writing Evaluation (AWE) tools for L2 learners (Ranalli, 2021). In particular, the design of the new generation of AWE tools has been increasingly aligned with findings from research in Second Language Acquisition and L2 writing to better accommodate the needs of English learners. This alignment includes increased identification of context-dependent writing errors, provision of automated feedback at different levels of granularity, and provision of multiple teaching and writing aids. As such, AWE holds great potential to meet the growing needs for online writing support in the current global pandemic and post-pandemic educational settings where "online teaching and learning tools have shifted from being an instructional asset to a necessity" (Limpo et al., 2020, p. 1).

Questions remain, however, to what extent L2 learners would engage with and benefit from AWE to improve their writing ability. Previous AWE research has primarily focused on the impact of using AWE on the quality of text produced (i.e., score and error rate) or the output of feedback processing (i.e., revision responses). This line of research has yielded inconclusive findings (Stevenson, 2016; Stevenson & Phakiti, 2014) on the efficacy of AWE in L2 contexts. With such a predominance of an outcome-based approach in research, how learners engage with and benefit from AWE feedback in the process of essay

revision appears as a black box (Stevenson & Phakiti, 2014; Storch, 2018; Warschauer & Ware, 2006). Among previous research, only six recent case studies (El Ebyary & Windeatt, 2019; Koltovskaia, 2020; Ranalli, 2021; Tian & Zhou, 2020; Zhang, 2020; Zhang & Hyland, 2018) have investigated how L2 learners engage with automated feedback. These studies have provided valuable (albeit limited) initial insights into why some learners benefit from using AWE while others do not, partly because of the very small sample size and the descriptive nature of these qualitative studies. Researchers (e.g., Stevenson & Phakiti, 2019; Storch, 2018) have been calling for more theoretically grounded and methodologically robust research to deepen our understanding of L2 learners' engagement with automated feedback.

In response to the call for research, this study investigated how 24 Chinese EFL learners engaged with feedback from a new AWE system and how the explicitness and accuracy of such feedback affected their engagement. Theoretically, we drew upon L2 writing research to conceptualise learner engagement with automated feedback. Methodologically, we combined eye-tracking with stimulated recall to collect and triangulate data for a more fine-grained analysis of how and why they engaged with automated feedback in a certain way.

Conceptualising Learner Engagement with Automated Feedback

Automated feedback is designed to emulate teacher-provided written corrective feedback; therefore, it would be beneficial for AWE research to align itself more closely with research on L2 writing (Stevenson, 2016). This study uses the conceptual framework that integrates the key processing stages of written corrective feedback proposed by Bitchener and Storch (2016, p. 20) and further elaborated in Bitchener (2017). Gass's (1997) model, which encompasses a full perspective of L2 learning processes involved in the use of feedback, is the basis of this conceptual framework. Learner engagement with feedback is defined as how learners process the feedback they received at five stages, including (a) Attention and Noticing of Input, (b) Comprehended Input, (c) Intake, (d) Integration, and (e) Output. Specifically, Attention and Noticing of Input (Stage 1) focuses on whether and to what extent a learner attends to the feedback provided. This stage forms the cornerstone of the entire feedback processing because it is a learner's conscious attention to the feedback that enables them to notice the gap between their language use and the target rule of L2 (Gass, 1997; Schmidt, 2001). Comprehended Input (Stage 2) is where feedback explicitness and language proficiency may determine a learner's ability to comprehend the feedback. At the Intake (Stage 3) and Integration (Stage 4) stages, learners match the feedback with their existing knowledge by forming and testing a hypothesis about what is acceptable and not in the L2. A rejected hypothesis leads to the formation of another hypothesis; a confirmed hypothesis helps strengthen the learner's prior knowledge or confirm a new use of the linguistic item. Stage 5, Output, is an overt manifestation of such processing reflected in the learner's revised text. If the revised text has been confirmed as an accurate use of the L2, the initial feedback episode is complete. If the revised text demonstrates an inaccurate use of the L2, the learner may need to restart the whole feedback processing based on further feedback, thus starting another episode of feedback processing. Compared to Ellis's (2010) tripartite conceptualisation of engagement (i.e., cognitive, behavioural, and emotional engagement), this framework explicates five micro key stages from processing feedback to producing revisions. It was adopted in the present study for a more nuanced understanding of L2 learners' engagement with automated feedback. In doing so, this study may also help advance research on automated feedback engagement by offering a new theoretical framework.

Informed by the framework (Bitchener, 2017; Bitchener & Storch, 2016), we conducted a critical review of previous studies of L2 learners' use of automated feedback by mapping them onto the five key stages of feedback processing. Based on the review, we defined the construct of learner engagement with automated feedback and formulated the research questions for the present study. The following sections present our critical review and research questions.

Stages of L2 Learners' Processing of Automated Feedback

Most of the previous research has investigated the output of automated feedback processing (Stage 5) in terms of learners' revision responses, such as feedback uptake, and type and quality of revisions. This line of inquiry has produced mixed results. Some researchers reported a low uptake rate of feedback from Pigai (11.5% in Bai & Hu, 2017; 24.2% in Tian & Zhou, 2020). Other researchers reported much higher rates of uptake of feedback from My Access (50% in Dikli, 2010) and Criterion (about 50% in Chapelle et al., 2015; 73% in Lavolette et al., 2015). It was found that learners' revisions mainly involved the correction of surface-level features such as spelling and capitalisation (e.g., Bai & Hu, 2017; Chen & Cheng, 2008; Lavolette et al., 2015; Zhang, 2020), largely due to the nature of automated feedback which mainly indicates surface-level errors (Stevenson, 2016). Research on revision quality has shown a range of success rates (i.e., the percentage of good/successful revisions), from 60% of Pigai-based revisions (Bai & Hu, 2017), to about 60% to 70% of Criterion-based revisions (Chapelle et al., 2015; Ranalli et al., 2017), and 57.2% to 85% of Grammarly-based revisions (Guo et al., 2021; Koltovskaia, 2020).

Only six recent case studies thus far have directly examined the other processing stages of automated feedback. Table 1 provides a summary of these studies.

Table 1

Summary of the Six Case Studies Examining Various Stages of Feedback Processing

	Number of participants	**AWE system**	**Stage of feedback processing**
El Ebyary & Windeatt (2019)	4	*Criterion*	Attention and Noticing of Input
Ranalli (2021)	6	*Grammarly*	Attention and Noticing of Input
Koltovskaia (2020)	2	*Grammarly*	Intake and Integration
Tian & Zhou (2020)	5	*Pigai*	Intake and Integration
Zhang (2020)	3	*Pigai*	Intake and Integration
Zhang & Hyland (2018)	2	*Pigai*	Intake and Integration

Among the six studies, two studies have investigated how learners attended to automated feedback at the Attention and Noticing of Input stage (Stage 1). El Ebyary and Windeatt (2019) combined eye-tracking with stimulated recall to examine which feedback the four L2 learners looked at first and for how long. It was found that all learners paid more attention to Criterion feedback on grammar rather than those on organisation and development, usage, mechanics, and style. Ranalli (2021) investigated six L2 learners' engagement time with Grammarly feedback utilising a combination of eye-tracking, stimulated recall, and interviews. In this paper, learners' average engagement time with feedback was calculated based on the "elapsed time from the moment each participant began reviewing Grammarly's first flagging until the final flagging has been resolved" (Ranalli, 2021, p. 7). The results showed that the average time learners spent on each Grammarly flagging ranged from 11 seconds to 19 seconds, which varied with their trust in and perceived value of such feedback.

The other four studies explored how L2 learners processed automated feedback at the Intake and Integration stages (Stages 3 & 4). Following Ellis's (2010) framework, they investigated learners' cognitive, behavioural, and emotional engagement mainly based on interview data and screen recordings. The three studies of learners' engagement with Pigai feedback (Tian & Zhou, 2020; Zhang, 2020; Zhang & Hyland, 2018) revealed that learners expended more effort in understanding the underlying rule of Pigai's indirect feedback on synonyms and collocations by drawing on their previous linguistic knowledge and seeking additional assistance (e.g., consulting a dictionary, searching online, or asking for

peer assistance). In a similar vein, Koltovskaia (2020) found that the learner with higher English proficiency drew on their linguistic knowledge and consulted the Internet to form and test the hypotheses of the underlying rules of Grammarly's feedback. However, the less proficient learner was cognitively overwhelmed and emotionally discouraged from seeking extra assistance to process the feedback, and thus defaulted to blind acceptance.

Previous studies have suggested feedback explicitness (Ranalli et al., 2017; Tian & Zhou, 2020; Zhang, 2020) as an important factor that may influence how L2 learners process automated feedback. Ranalli (2018) extended previous research by examining 82 L2 learners' error correction performance based on mock-ups of Criterion's generic and specific feedback. The results showed that, compared to generic feedback, specific feedback was considered clearer and more useful and led to more successful revisions. It was therefore hypothesised that feedback explicitness may be a "determining" factor affecting L2 learners' use of automated feedback (Ranalli, 2018, p. 668) and may explain the difficulty they faced in engaging with some Criterion feedback. However, this hypothesis needs to be tested in further research because the mock-up tasks may not reflect learners' real-time processing of automated feedback. Another strong explanatory factor is feedback accuracy. Several studies (Bai & Hu, 2017; Chapelle et al., 2015; Koltovskaia, 2020; Lavolette et al., 2015) have suggested that learners' uncertainty about AWE feedback accuracy may have led them to reject the feedback. Such an argument was supported by Guo et al. (2021), who found that the accuracy of Grammarly feedback was significantly related to users' feedback uptake and revision quality. This study was among the first to assess the impact of feedback accuracy statistically and has shed fresh light on the relationship between feedback accuracy and engagement. However, since it only focused on learners' revision responses (i.e., the Output stage), it remains unknown how much the feedback accuracy may affect their feedback processing at the other stages.

Based on the critical review of previous AWE research on feedback processing, we conceptualised learner engagement with automated feedback as composed of three key interrelated elements:

- **Attention Allocation** (at Stage 1, Attention and Noticing of Input) concerns whether and how much a learner allocates attention to the automated feedback provided.
- **Cognitive Effort Expenditure** (at Stage 3, Intake, and Stage 4, Integration) refers to the cognitive capacity allocated by the learner to form or test a hypothesis of the underlying rule(s) of the automated feedback.
- **Revision Response** (at Stage 5, Output) is the manifestation of a learner's automated feedback processing. It is reflected in the learner's feedback uptake (i.e., the incorporation of feedback in revision), revision type (i.e., surface or content revisions), and revision quality (i.e., the percentage of good/successful revisions).

Our conceptualisation did not focus on Stage 2, Comprehended Input, because it is more closely related to the underlying factors that affect the processing of AWE feedback. Based on this conceptualisation, three research gaps can be identified. First, none of the previous research has investigated all three elements to provide a comprehensive picture of how L2 learners engage with automated feedback. Second, the impact of feedback explicitness and accuracy on L2 learners' feedback engagement remains understudied; in particular, no research has investigated the impact of both feedback explicitness and accuracy (Ranalli, 2021). Third, existing studies have exclusively focused on major AWE systems such as Pigai, Criterion, and Grammarly. To the best of our knowledge, no study thus far has examined learner engagement with Write & Improve with Cambridge, a new AWE system developed by a major test provider, Cambridge University Press & Assessment, targeting L2 English learners. Such investigation is urgently needed as the findings may provide a timely contribution to the current understanding of learner engagement with automated feedback and indicate further directions for the development of AWE as educational technology. The present study aims to bridge these gaps by investigating all three elements of EFL learners' engagement with a new AWE system and by examining the impact of feedback explicitness and accuracy on such engagement. Two research questions (RQs) were formulated in this study:

1. How do EFL learners engage with automated feedback in terms of their Attention Allocation, Cognitive Effort Expenditure, and Revision Responses?

2. To what extent do feedback explicitness and accuracy affect EFL learners' engagement with automated feedback?

Methodology

Write & Improve with Cambridge

Write & Improve with Cambridge, henceforth W&I, is a free online English learning and test preparation platform for L2 learners. Learners can practise English writing on a wide range of topics at various proficiency levels using the system's preloaded prompts or tailor-made tasks created by the teacher, and receive automated feedback on their writing. W&I provides three types of diagnostic feedback to learners. Direct Word-level Feedback (DWF) provides the most detailed information, including error location, explanation, and suggested correction (Figure 1). Indirect Word-level Feedback (IWF) is less explicit as it only flags a word/phrase and briefly explains the error (Figure 2). Indirect Sentence-level Feedback (ISF) indicates the quality of each sentence by highlighting them in different colour shades, with the darker colour background suggesting more problematic sentences (Figure 1). ISF is the least explicit as it provides no information about "the nature and cause of the error and what is needed for an accurate modification" (Bitchener, 2017, p. 133).

Figure 1

Screenshot of DWF by W&I

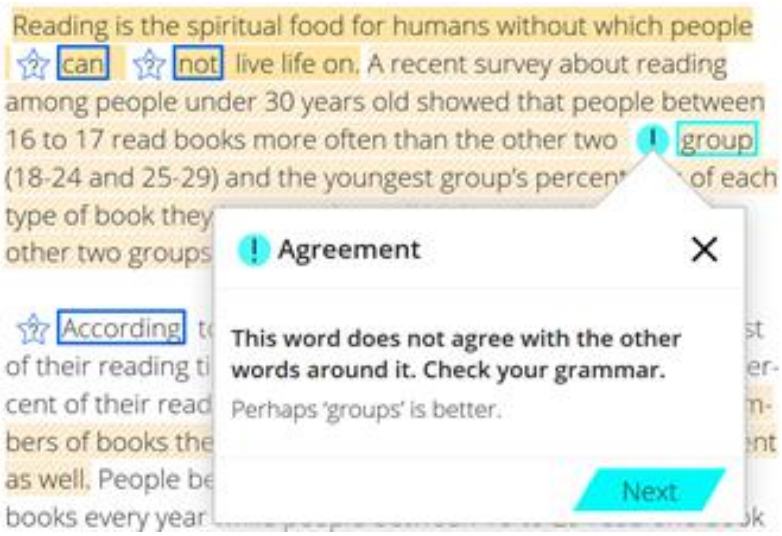
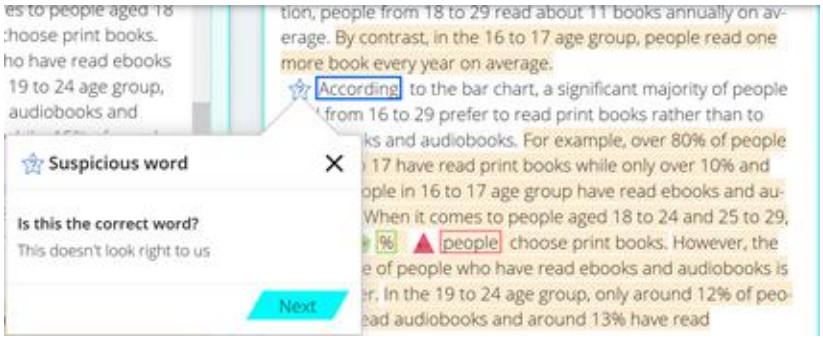


Figure 2

Screenshot of IWF by W&I



The provision of these three types of feedback distinguishes W&I from other major AWE systems such as Criterion, My Access, Grammarly, WriteToLearn, and Pigai. As summarised in Table 2, instead of generating direct and/or indirect feedback, W&I treats them as two ends of a continuum of feedback with varying degrees of explicitness. In particular, the system indicates the presence of a sentence-level error without providing any additional information (i.e., ISF), whereas none of the other systems do so due to their technological capacities (Ranalli, 2018).

Table 2

Types of Automated Feedback from Different AWE Systems

AWE system	Type of automated feedback
Criterion	Direct and indirect feedback (Ranalli, 2018)
My Access	Direct feedback only (Hoang & Kunnan, 2016)
WriteToLearn	Direct feedback only (Liu & Kunnan, 2016)
Grammarly	Direct feedback only (Koltovskaia, 2020; Ranalli, 2021)
Pigai	Direct and indirect feedback (Bai & Hu, 2017)
W&I	Direct and indirect feedback with different levels of explicitness

Participants and Setting

The study took place at a major university in Southwest China. An entire class of 25 English majors aged 18 to 20 were recruited. Among the participants, there were 22 females and three males. All participants spoke Chinese as a first language, and they had approximately 10.2 years of English learning experience on average. Their English proficiency level ranged from B1 to C1, as evident by the W&I grades of their writing. All participants had prior experience using Pigai, a locally designed and most widely used AWE system in China (Bai & Hu, 2017), and reported willingness to try out different AWE systems for the improvement of their English writing.

Data Collection

The data was collected from participants' engagement with feedback on their writing from two graph-based integrated writing prompts (see Appendix A, Prompt 1 "Book Reading and Types of Book" and Prompt 2 "Trust in News Sources"). Both prompts were retired IELTS Academic Writing Tasks preloaded in W&I. These prompts were chosen because participants reported a lack of familiarity with such tasks compared to their writing practice in response to other types of reading/writing integrated tasks.

For each prompt, participants were instructed to revise their essays through W&I on a laptop installed with Tobii Studio (version 3.4.8; Tobii Technology, 2017). A portable Tobii-X2-60 eye tracker was attached to the bottom frame of the laptop. Each participant started their essay revision with the first author on a scheduled day of data collection. Before the revision process, the researcher calibrated each participant's eye movements based on a 9-point calibration. One participant was excluded because the calibration failed with her eye movements, and the study was thus left with 24 participants. Each participant then revised their essay through W&I for 30 minutes and was allowed to submit multiple drafts; they were not allowed to access any external resources (e.g., dictionaries and phones). During this process, their eye movements were tracked; simultaneously, the researcher watched their eye movements through the live view function of Tobii Studio on another computer and took notes about their feedback viewing and essay revision behaviours as shown in the live view. In addition, participants' emotional reactions (e.g., a sigh) that the researcher observed at specific time points were also noted. The field notes were utilised for more targeted and efficient data elicitation in the stimulated recall.

Stimulated recall interviews were conducted with each participant immediately after the eye-tracking. The interview was prompted by each participant's eye-gaze replay in Tobii Studio and the researcher's field notes. To facilitate participants' understanding of the replay of their eye movements, the researcher first delivered an introduction to their eye movement data. It was explained in plain language that circles represented their eye fixations, and lines represented switching between successive eye fixations; the larger the circle, the longer they dwelled on a specific area. During the interview, particular attention was paid to elicit data about participants' online processing of the three types of W&I feedback. Participants were asked to elaborate on whether they understood the underlying rule of each instance of the feedback and what effort they made to achieve such understanding. The interview questions are reported in Appendix B. Participants were also encouraged to ask the researcher to pause the replay at any time and comment on anything that they saw as important or interesting to discuss. For ease of communication, all interviews were conducted in Chinese and audio recorded.

We collected 48 sets of eye gazeplot recordings together with the corresponding stimulated recall interviews (i.e., 24 sets of data for each writing prompt). Each eye gazeplot recording lasted for 30 minutes, and the interview lasted for about 45 minutes for each writing task on average. The three types of automated feedback (i.e., DWF, IWF, and ISF) that participants received during the 30-minute eye-tracking session were exported from W&I. At the end of the study, participants were required to write a 500-word journal in English to reflect on the impact using W&I had on their writing.

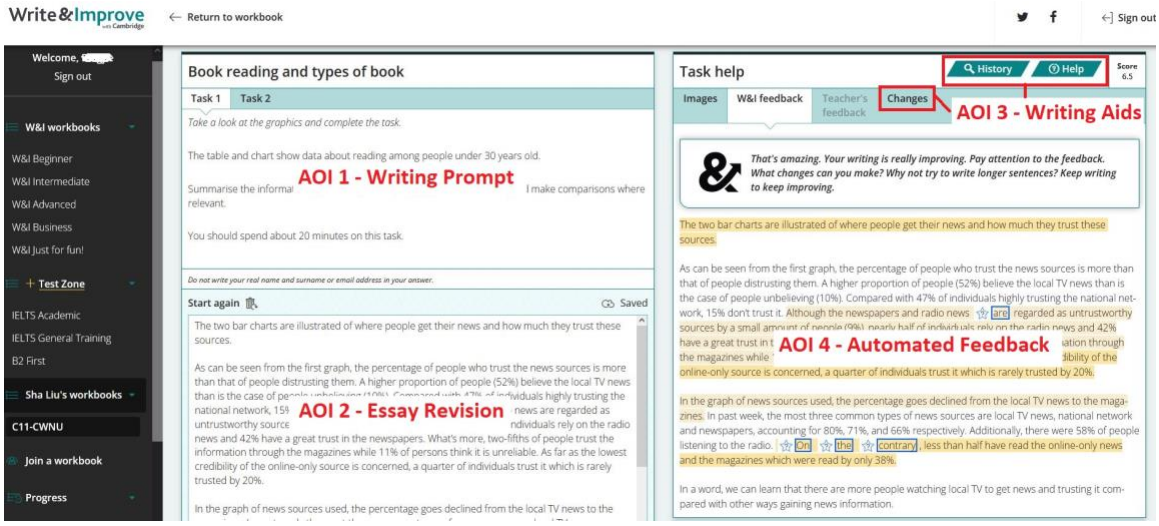
Data Analysis

Eye Gazeplot Recordings

The video recordings of participants' eye gazeplots and their revision processes in multiple drafts were manually annotated and coded in NVivo 12 (QSR International Pty Ltd., 2018). In an eye gazeplot recording, Tobii Studio video-recorded participants' every activity on the computer screen (i.e., keystroke, mouse click, and pause) simultaneously with their eye movements (i.e., fixations and saccades). The manual annotation was based on a coding scheme developed and pilot tested by the researcher (Appendix C). The coding scheme was designed in accordance with the main Areas of Interests (AOIs) of this study (Figure 3) and their sub-AOIs.

Figure 3

Four Main AOIs in This Study



Within AOI 4, there was not a single "fixed" stimulus at any given point in time because every participant was presented with different feedback. Therefore, the time participants spent reading each individual instance of the three types of feedback (i.e., DWF, IWF, and ISF) was coded manually. Then, their total fixation duration on a specific type of feedback was calculated based on its coding coverage on NVivo. For example, the coding of a participant's reading of DWF covered 7.39% of the 30-minute gazeplot recording (Figure 4); therefore, the participant's total fixation duration on DWF was 133.02 seconds (i.e., 7.39% * 30 minutes * 60 seconds). Total fixation duration refers to the duration of all fixations a participant made and is thus an important indicator of how much time a participant spends on an area of interest (Yu et al., 2017). It has been proven to be a valid measure to infer conscious effort of reading and mindful cognitive processing of feedback (e.g., Bolzer et al., 2015) and therefore used in this study to indicate participants' attention allocation to feedback.

To ensure the quality of data analysis, 33% of randomly selected eye gazeplot recordings with corresponding interview transcripts (see Stimulated Recall Interviews) were double coded by a highly proficient EFL colleague and the first author. Intercoder reliability was high, with an average Cohen's kappa of .91 for the analysis of eye movements and .89 for the analysis of interviews. Any disagreement in coding was resolved through discussion until a consensus was reached. The researcher then carried out an analysis of the remaining data.

Figure 4

Screenshot of the Coding Coverage of a Participant's Reading of DWF

Stimulated Recall Interviews

Participants' interview responses in Chinese were first transcribed verbatim and sent to the corresponding participant for a validity check. Then, the confirmed transcripts were directly analysed in Chinese to avoid distorting the original voice of the participants. The transcripts were analysed in NVivo 12 via a hybrid approach of deductive and inductive thematic analysis (Fereday & Muir-Cochrane, 2006). They were initially examined for participants' verbalisations of what cognitive effort they expended to understand each piece of feedback. An instance of cognitive effort expenditure was coded when a participant reported forming or testing the feedback's underlying rule; no occurrence of such effort was coded if a participant reported directly adopting the feedback without any attempt to understand the underlying rule or making a revision by guessing the correct form (see Appendix D). The total number of coded instances of cognitive effort expenditure in each feedback type were then summed. Any recurring and emergent themes related to the impact of feedback explicitness and accuracy on participants' feedback engagement were then identified by reiteratively examining the transcripts.

Reflective Journals

Participants' reflective journals were double coded by the EFL colleague mentioned earlier and the first author based on Fereday and Muir-Cochrane (2006). They were reiteratively examined to identify themes related to participants' perceived impact of using W&I feedback; the final codes included Motivation, Noticing the Errors, and the Development of Independent Revision Skills. Intercoder reliability was very high, with an average Cohen's kappa of .99.

Revision Responses

Participants' feedback uptake was operationalised as the amount of feedback that was accepted and taken up. An instance of feedback was coded as take-up if participants made any revision based on the feedback as revealed in the gazeplot recording. The total cases of adopted feedback in each feedback type were then summed. There were two types of revisions: surface and content. Surface revisions were made to mechanics, grammar, or the order of sentences which did not affect the meaning of the text (i.e., meaning preserving). Content revisions changed the informational content at a local or global level of the text (i.e., meaning changing). The coding scheme was developed based on Barkaoui (2016, p. 326) but was modified to fit the data from the study (see Appendix E). Revision quality was classified into good, neutral, and bad revisions based on the coding scheme by Bai and Hu (2017, p. 72): good revisions "correctly fix problems, clarify ideas or improve expressions"; neutral revisions "neither improve nor worsen the well-formed or ill-formed original text"; bad revisions "produce errors or degrade the quality of the original text" (see Appendix F). The classification of the type and quality of revisions was undertaken by an English native-speaking colleague with about ten years of L2 teaching experience and the first author. Intercoder reliability was excellent, with an average Cohen's kappa of .99 for both analyses. Based on the coding, the total number of content revisions and good revisions in each feedback type were then summed. The success rates of revisions were operationalised as the percentage of good revisions participants made.

Accuracy of W&I Feedback

Each instance of the three types of feedback (i.e., DWF, IWF, ISF) was coded as accurate or inaccurate by the same English native-speaking colleague and the first author. Intercoder reliability was good, with an average Cohen's kappa of 0.87. Any disagreements in coding were resolved through discussions until a consensus was reached. The finalised coding was then used to calculate the accuracy rate of each feedback type.

Statistical Analysis of the Impact of Feedback Explicitness and Accuracy

Wilcoxon Signed Rank Tests were conducted to compare differences in participants' Attention Allocation,

Cognitive Effort Expenditure, and Revision Responses. Nonparametric procedures were used because the data were not normally distributed. To control for the differences in feedback frequency, participants' total fixation duration and cognitive effort expenditure were divided by the number of each type of feedback they received; their revision responses (i.e., feedback uptake, revision type, and revision quality) were converted into percentages. According to Cohen's (1988) benchmarks of effect size *r*, values larger than .1, .3, and .5 were considered as small, medium, and large, respectively. The statistical findings were triangulated with participants' stimulated recall interviews and reflective journals.

Results

Table 3 provides an overview of the feedback that participants received on their writing. There was considerable variation in the amount of the three types of feedback. All participants received ISF, with 438 sentences highlighted. The DWF and IWF they received were relatively sparse. One participant (Participant #17) did not receive any DWF, and another (Participant #4) did not receive any IWF; in total, participants only received 150 instances of DWF and 109 instances of IWF. The accuracy of the three types of feedback also varied. The precision of IWF (55.05%) was markedly lower than those of DWF (83.33%) and ISF (94.06%); this was because IWF mostly flagged word choice errors, the detection of which has been found to be a thorny issue for AWE systems (Liu & Kunnan, 2016).

Table 3

Summary of W&I Feedback

		N	Total	*M*	*SD*	Accuracy
DWF	Accurate	23	125	5.43	4.82	83.33%
	Inaccurate	16	25	1.56	.81	
	Overall	23	150	6.52	4.75	
IWF	Accurate	20	60	3.00	1.414	55.05%
	Inaccurate	21	49	2.33	1.592	
	Overall	23	109	4.74	2.16	
ISF	Accurate	24	412	17.17	4.39	94.06%
	Inaccurate	14	26	1.86	1.17	
	Overall	24	438	18.25	4.39	

Note. *N* = The number of participants who received the feedback; Total = The total number of the instances of feedback participants received.

Participants' Engagement with W&I Feedback (RQ1)

Attention Allocation

Table 4 shows descriptive statistics of participants' total fixation duration on feedback. Participants allocated 16.14 seconds on average to read each instance of DWF, which is much shorter than 21.62 seconds for IWF and 39.59 seconds for ISF. They spent less time reading the accurate DWF, IWF, and ISF (*M* = 14.91, 18.74, 39.75, respectively) than the inaccurate feedback (*M* = 19.81, 28.84, 40.18, respectively). Overall, participants spent 32.39 seconds on average reading each W&I flagging.

Table 4

Descriptive Statistics of Participants' Total Fixation Duration (Seconds)

		Raw		Standardised	
		M	***SD***	***M***	***SD***
DWF	Accurate	62.52	43.68	14.91	10.36
	Inaccurate	27.72	31.25	19.81	21.13
	Overall	78.40	51.15	16.14	10.28
IWF	Accurate	48.67	31.58	18.74	11.36
	Inaccurate	53.73	35.31	28.84	27.77
	Overall	87.58	43.35	21.62	10.26
ISF	Accurate	634.90	318.47	39.75	24.49
	Inaccurate	77.63	85.06	40.18	37.19
	Overall	680.18	347.09	39.59	24.58
Total		846.16	361.35	32.39	18.58

Cognitive Effort Expenditure

Table 5 presents descriptive statistics of participants' cognitive effort expenditure in processing feedback. Participants made fewer attempts to understand DWF (M = .91) than IWF (M = 1.57) and ISF (M = 2.21). Less effort was expended to understand the accurate DWF, IWF, and ISF (M = 1.06, 1.46, 2.19, respectively) than the inaccurate feedback (M = 1.27, 1.80, 2.49, respectively). Overall, participants made 1.85 attempts on average to form or test the underlying rule of each W&I flagging.

Table 5

Descriptive Statistics of Participants' Cognitive Effort Expenditure (Counts)

		Raw		Standardised	
		M	***SD***	***M***	***SD***
DWF	Accurate	3.26	3.31	1.06	1.57
	Inaccurate	1.94	2.35	1.27	1.43
	Overall	4.61	4.74	.91	.98
IWF	Accurate	3.90	2.43	1.46	1.00
	Inaccurate	3.71	3.35	1.80	1.14
	Overall	6.83	4.65	1.57	.96
ISF	Accurate	35.54	13.98	2.19	1.00
	Inaccurate	5.07	4.53	2.49	1.25
	Overall	38.50	15.40	2.21	.96
Total		49.46	19.94	1.85	.92

Revision Responses

Table 6 provides a summary of participants' revision responses to feedback. As for feedback uptake, participants incorporated a much higher percentage of DWF (91.94%) than the indirect feedback (72.75% for IWF and 71.85% for ISF, respectively). They took up a higher percentage of accurate DWF, IWF, and ISF (93.66%, 75.42%, 71.62%, respectively) than the inaccurate feedback (81.25%, 63.65%, 68.81%, respectively). Overall, participants took up 75.12% of total W&I feedback in their revisions.

In terms of revision type, the percentage of content revisions in response to DWF (5.62%) was much lower than those of IWF (18.14%) and ISF (26.94%). Participants made more content revisions to inaccurate DWF and ISF (9.99% and 54.08%, respectively) than accurate feedback (3.44% and 26.01%, respectively). However, within IWF, more content revisions were made to the accurate (25.37%) than the inaccurate feedback (11.76%). Overall, only 22.59% of the revisions that participants made based on W&I feedback were related to the content of the text.

Turning to revision quality, participants' success rate of DWF-based revisions was twice as high (79.87%) as the IWF-based (39.62%) and ISF-based revisions (38.64%). They were more successful in error correction based on accurate DWF, IWF, and ISF (88.75%, 37.68%, 38.87%, respectively) than inaccurate feedback (39.70%, 8.82%, 23.17%, respectively). Overall, participants were only able to correct 44.92% of total errors flagged by W&I.

Table 6

Descriptive Statistics of Participants' Revision Responses (Counts)

		Feedback uptake			Content Revisions			Successful Revisions		
		M	***SD***	**%**	***M***	***SD***	**%**	***M***	***SD***	**%**
RDWF	Accurate	5.04	4.70	93.66	.22	.60	3.44	5.09	5.31	88.75
	Inaccurate	1.31	1.01	81.25	.06	.25	9.09	.56	.73	39.70
	Overall	5.96	4.63	91.94	.26	.62	5.62	5.48	5.41	79.87
RIWF	Accurate	2.20	1.40	75.42	.45	.69	25.37	2.00	1.72	37.68
	Inaccurate	1.43	1.17	63.65	.10	.30	11.76	.10	.30	8.82
	Overall	3.52	2.33	72.75	.48	.73	18.14	1.83	1.70	39.62
RISF	Accurate	12.00	3.20	71.62	8.46	3.20	26.01	13.04	5.24	38.87
	Inaccurate	1.29	1.07	68.81	1.07	1.27	54.08	.86	1.23	23.17
	Overall	12.79	3.24	71.85	9.08	3.09	26.94	13.54	5.67	38.64
Total		21.58	6.94	75.12	9.79	3.27	22.59	20.54	6.99	44.92

Note. RDWF = Revisions to DWF, RIWF = Revisions to IWF, RISF = Revisions to ISF.

The Impact of Feedback Explicitness and Accuracy on Participants' Feedback Engagement (RQ2)

Quantitative Findings

Impact of Feedback Explicitness

The Friedman Tests showed significant differences in participants' engagement with the three types of feedback (overall $p < .001$). We then conducted post-hoc Wilcoxon Singed Rank Tests. Table 7 summarises the test results.

Table 7

Wilcoxon Signed Rank Test Results by Feedback Type

		N	***z***	***p*.adj**	***r***
Total Fixation Duration (Seconds)	DWF - IWF	22	-2.484	.038*	-.53
	IWF - ISF	23	-3.406	.002**	-.71
	DWF - ISF	23	-3.924	.002**	-.82
Cognitive Effort Expenditure (Counts)	DWF - IWF	22	-3.911	.002**	-.83
	IWF - ISF	23	-4.076	.000**	-.85
	DWF- ISF	23	-4.167	.000**	-.87
Feedback Uptake (Percentage)	RDWF - RIWF	22	-1.762	.234	-.38
	RIWF - RISF	23	-.1.445	.445	-.30
	RDWF- RISF	23	-3.718	.000**	-.78
Content Revisions (Percentage)	RDWF - RIWF	20	-1.780	.225	-.38
	RIWF - RISF	21	- 2.038	.125	-.42
	RDWF - RISF	23	-4.198	.000**	-.88
Good Revisions (Percentage)	RDWF - RIWF	22	-3.885	.000**	-.83
	RIWF - RISF	23	-.061	1.000	-.01
	RDWF - RISF	23	-4.045	.000**	-.84

Notes. N = The number of participants who received and responded to the feedback; $r = z/\sqrt{N}$; *p.adj* = Adjusted p-value based on Bonferroni corrections. *$p < .05$, **$p < .01$

In terms of participants' attention to feedback, the tests revealed statistically significant differences in all three pairs of comparison, including a) between DWF (Median = 15.84) and IWF (Median = 21.06, z = -2.484, $p < .05$, r = -.53), b) between IWF and ISF (Median = 31.45, z = -3.406, $p < .01$, r = -.71), and c) between DWF and ISF(z = -3.924, $p < .01$, r = -.82). The effect sizes of all these differences were large.

With respect to cognitive effort expenditure, statistically significant differences were also found in all three pairs of comparison, including a) between DWF (Median = .50) and IWF (Median = 1.25, z = -3.911, $p < .01$, r = -.83), b) between IWF and ISF (Median = 2.10, z = -4.076, $p < .01$, r = -.85), and c) between DWF and ISF (z = -4.167, $p < .01$, r = -.87). The effect sizes of all these differences were large.

The tests revealed statistically significant differences between participants' uptake of DWF (Median = 100%) and those of ISF (Median = 71.53%, z = -3.718, $p < .01$), with a large effect size (r = -.78). No statistically significant difference was found in the DWF versus IWF comparison (z = -1.762, p = .234) and IWF versus ISF comparison (z = -1.445, p = .445).

As for revision type, a statistically significant difference was found between the percentage of content revisions in response to DWF (Median = 0%) and ISF (Median = 25.68%, z = -4.198, $p < .01$), with a large effect size (r = -.88). No such differences were found in the DWF versus IWF comparison (z = -1.780, p = .225) and the IWF versus ISF comparison (z = -2.038, p = .125).

In terms of revision quality, the tests found statistically significant differences between the percentage of successful revisions in response to DWF (Median = 80.00%) and that to IWF (Median = 42.86%, z = -3.885, $p < .01$, r = -.83) and to ISF (Median = 39.23%, z = -4.045, $p < .01$, r = -.84). The effect sizes of these two differences were large. No statistically significant difference was found between the IWF and ISF comparison (z = -.061, p = 1.000).

In sum, participants spent more time reading, expended more cognitive effort processing, and made more content revisions to the indirect than direct feedback. However, a lower percentage of indirect feedback was taken up, and the revisions participants made based on such feedback were less successful. Such results suggested that feedback explicitness plays a crucial role in mediating participants' feedback engagement.

Impact of Feedback Accuracy

The Friedman tests showed significant differences in participants' attention to, cognitive effort expenditure in, and uptake of accurate and inaccurate W&I feedback (overall $p < .05$) while showing non-significant differences in their type and quality of revisions to such feedback (p = .092 and .098, respectively). We then conducted post-hoc Wilcoxon Singed Rank Tests. Table 8 summarises the test results. No statistically significant difference was observed in any of the 15 pairs of comparisons. These results indicate that the accuracy of W&I feedback had no impact on participants' feedback engagement.

Table 8

Wilcoxon Signed Rank Test Results by Feedback Accuracy

		N	***z***	***p*.adj**	***r***
Total Fixation Duration (Seconds)	Accurate - Inaccurate DWF	16	-.052	1.000	-.01
	Accurate - Inaccurate IWF	18	-1.808	.212	-.43
	Accurate - Inaccurate ISF	14	-1.224	.663	-.33
Cognitive Effort Expenditure (Counts)	Accurate - Inaccurate DWF	16	-.251	1.000	-.06
	Accurate - Inaccurate IWF	18	-1.108	.803	-.26
	Accurate - Inaccurate ISF	14	-.534	1.000	-.14
Feedback Uptake (Percentage)	Accurate - Inaccurate DWF	16	-1.214	.675	-.30
	Accurate - Inaccurate IWF	18	-.981	.979	-.23
	Accurate - Inaccurate ISF	14	-.220	1.000	-.06
Content Revisions (Percentage)	Accurate - Inaccurate DWF	11	-1.000	.952	-.30
	Accurate - Inaccurate IWF	14	-.813	1.000	-.22
	Accurate - Inaccurate ISF	10	-1.785	.223	-.56
Successful Revisions (Percentage)	Accurate - Inaccurate DWF	11	-2.231	.077	-.67
	Accurate - Inaccurate IWF	17	-2.216	.080	-.41
	Accurate - Inaccurate ISF	10	-1.955	.152	-.62

Note. N = The number of participants who received and responded to the feedback; $r = z/\sqrt{N}$; *p.adj* = Adjusted p-value based on Bonferroni corrections.

Qualitative Findings

Feedback explicitness was found to be determinative of participants' strategic engagement sequence with and perception of the learning potential of the three types of W&I feedback. All participants started with DWF because they could easily figure out the errors and correct them by simply accepting the suggested forms. However, while accepting the ease of understanding and error correction as mood-boosting, participants suspected that such superficial engagement would not actually improve their ability to correct errors in the future:

> I simply accepted its suggestions and after resubmission, all the error flagging just disappeared. I was happy but also wondered whether I would be able to correct such errors next time. (P#15_P2)

Participants then moved on to IWF, which was considered more difficult to engage with because of the need to self-edit errors. IWF flagged participants' frequent but erroneous use of words that they assumed to be synonyms at a relatively low accuracy rate (55.05% in Table 3). In response, participants initiated more extensive engagement by drawing upon their previous linguistic knowledge and experience to figure out the underlying rule for each instance of feedback and evaluate its accuracy. For example, W&I flagged the phrase "according to" as problematic, the use of which was deemed acceptable by both human raters. When asked why they frequently revisited the feedback, participants repeatedly expressed their uncertainty about its accuracy as their previous use of "according to" had never been flagged by their teachers. Despite their uncertainty, all participants but two (P#6 & P#23) revised this phrase by either changing its spelling or substituting it with similar expressions such as "based on" and "as shown in," all of which were unsuccessful revisions (see Appendix F). This example clearly demonstrates that the accuracy of IWF had affected participants' feedback engagement in terms of their attention allocation, cognitive effort expenditure, and revision responses.

In addition to confirming the impact of feedback accuracy, participants' engagement with IWF has also revealed the difficult nature of IWF-based revisions and shown that it often resulted in a guessing game where participants kept substituting the flagged word with similar expressions that they could think of:

> I replaced "according to" with "based on" because I was more certain about the latter, although I had no idea whether there was any difference between them at all. (P#21_P2)

The difficulty of revision notwithstanding, all participants reported learning gains from engaging with IWF in terms of their increased awareness of the need to improve the accuracy of word choices both during the interviews and in their reflective journals:

> Before using the system, I used these words interchangeably for the lexical diversity of my writing. W&I made me realise that there may be some differences in these words. Now I will always look them up for a more accurate use. (P#9_P1)

Participants addressed ISF last because they found it highly challenging to self-diagnose and self-edit the errors in the highlighted sentences. To revise the highlighted sentences, participants examined all aspects of writing, including mechanics, grammar, accuracy of content, and even organisation, which could explain why this type of feedback resulted in a greater proportion of content revisions. In addition, they engaged all possible sources of information available in the user interface of W&I, including checking the writing prompt, monitoring changes they had made, and clicking to open the 'Help' function. Such extensive engagement, however, did not transfer to a high percentage of feedback uptake or success rate for revision, which could be because participants had insufficient linguistic knowledge to identify the nature of the errors and make self-corrections. For example, Participant #17 reflected on how their English language proficiency prevented them from successfully revising a highlighted sentence after 14 revision attempts:

> Because of my poor English ability, I just could not figure out what was wrong. Without knowing what the errors were, I could only take guesses and keep trying out different revisions. This process was so frustrating! Could you tell me what was wrong with this sentence? (P#17_P1)

Despite being cognitively and emotionally overwhelmed, all participants reported a strong preference for this type of feedback. Such a preference was, as revealed in all participants' reflections, particularly because of the realisation that ISF has helped them develop independent self-editing skills. It was in sharp

contrast to their feeling of being 'spoon-fed' in their previous use of Pigai, which was constantly mentioned in the interviews and in the reflective journals:

> Because the feedback of this system isn't specific enough, I have to think where the problem is by my own. I think this system helps me develop the habit which I have to solve problems by my own first. I have used to being given answers by teacher or the system [Pigai], but now, I will try to solve the problems by myself. So, this system has changed my studying habit. (P#11_Reflective Journal)

Discussion and Conclusion

This study deployed eye-tracking, stimulated recall and reflective journals to investigate L2 learners' engagement with automated feedback in terms of their attention allocation, cognitive effort expenditure, and revision responses, and the impact of feedback explicitness and accuracy on such engagement. The findings indicated participants' overall active engagement with W&I feedback in the process of essay revision as demonstrated by their a) duration of attention allocation (i.e., 33 seconds on average), which was much longer than the 11 to 19 second duration per Grammarly flagging in Ranalli (2021), b) continuous efforts in understanding the underlying rules of W&I feedback (i.e., two attempts on each flagging), and c) higher rate of feedback uptake (i.e., 75.12%) than those reported in previous studies, which ranged from 11.5% of Pigai feedback (Bai & Hu, 2017), 50% of My Access feedback (Dikli, 2010), and 73% of Criterion feedback (Lavolette et al., 2015).

With respect to revision type, participants only made a small percentage (22.59%) of content-related revisions. This finding suggests that W&I, like other major AWE systems, may largely induce surface revisions because of its limited technological capacity to evaluate and provide content-sensitive feedback. In terms of revision quality, participants' overall success rate (44.92%) was much lower than 57.2% of Grammarly-based revisions (Koltovskaia, 2020), 60% of Pigai-based revisions (Bai & Hu, 2017), and 60%–70% of Criterion-based revisions (Chapelle et al., 2015; Ranalli et al., 2017, respectively). As noted earlier, this overall low success rate was due to participants' failure to address the two forms of indirect feedback from W&I (i.e., IWF and ISF), even though they spent significantly more time and expended more cognitive effort in dealing with such feedback. These findings confirmed feedback explicitness as a determining factor affecting L2 learners' ability to engage with and benefit from automated feedback (Ranalli, 2018; Tian & Zhou, 2020; Zhang, 2020).

An intriguing finding of this study is how the direct and indirect feedback from W&I stimulated participants' different levels of engagement. DWF mostly stimulated superficial engagement; because of this, participants perceived little potential of such feedback to promote their L2 learning (e.g., improving their ability to correct errors in the future). This finding may explain why some learners in Ranalli (2021) tended to use the direct feedback from Grammarly only for proofreading purposes. The indirect feedback (i.e., IWF and ISF) prompted more extensive engagement and deeper processing. IWF helped participants identify areas for improvement (e.g., the erroneous use of synonyms) and increase their awareness of ensuring accuracy of word choices. ISF led to the change of learners' habit of relying on automated feedback and greatly motivated them to keep revising the highlighted sentences for the opportunity to develop their self-monitoring/editing skills. As such, learners ascribed much value to the two forms of indirect feedback, which resonates with Zhang and Hyland's (2018) conclusion that their participants saw more potential in Pigai's indirect feedback on synonyms and collocation. However, ISF could be highly challenging for L2 learners who may not have a big store of linguistic knowledge to draw on to respond successfully to such feedback (see also Koltovskaia, 2020; Zhang, 2020; Zhang & Hyland, 2018).

The quantitative data indicated that the accuracy of W&I feedback did not significantly affect participants' engagement with W&I feedback. This finding is at odds with previous studies, which found that the accuracy of AWE feedback may significantly affect L2 learners' ability to correct errors (Bai & Hu, 2017; Guo et al., 2021; Lavolette et al., 2015). Our qualitative data, however, showed that all participants

reported that they experienced more challenges in processing the inaccurate rather than the accurate IWF.

These findings have important implications for the further development of AWE and its use to promote successful engagement and optimum learning. AWE tools need to be constantly updated to improve their feedback accuracy; otherwise, error-prone feedback may cause participants to make inaccurate amendments to their writing (as with the case of "according to" in this study). In addition, it is suggested that AWE systems should keep a sensitive balance in the amount and type of direct and indirect feedback and, ideally, allow learners to decide how much, at what level of explicitness, and in what sequence they should be presented the feedback to better tailor to their individual linguistic needs, cognitive capacity, and mental wellbeing. There is also the need to enable learner-learner interaction and learner-teacher interaction in the feedback loop. For example, a dashboard can be created where learners can share their difficulties in revisions anonymously for peer and teacher support when it is beyond their language ability to process the feedback. Such peer-to-peer and learner-teacher interactions, according to Nassaji and Kartchava (2017), could lead to learners' active participation in the stages of processing feedback and engaging actively in higher levels of L2 learning more generally. Furthermore, further supplementary and scaffolding information, which is developmentally appropriate to individual learners, should be provided alongside ISF. For example, the systems may include the name of the error type (e.g., a run-on sentence), a glossary that defines the error type in both English and the learner's native language, and/or a demonstration of how to correct a similar error in a different context.

Teachers also have important roles to play in order to make the best use of AWE systems. We suggest that teachers should not leave learners to the systems alone because of "the vulnerability [learners may] experience when asked to rely on tools" (Ranalli, 2021, p. 39). Instead, they should help learners confirm the accuracy of AWE feedback (see also Koltovskaia, 2020) and monitor learners' engagement with such feedback to provide timely scaffolding. This is especially important for ISF because of learners' provisional grasp of the L2. This support may be in the form of tutorials and/or classroom activities (e.g., group presentations) to address the difficulty in feedback processing. For example, discussions can be held on why a specific sentence has been flagged as erroneous and how to address it at regular tutorial meetings.

The study has several limitations. The present study adopted an experimental design where participants were required to complete their revisions in 30 minutes without access to additional resources such as dictionaries. Therefore, further research may be needed to corroborate the findings in naturalistic settings. All participants were English majors who appeared to be highly motivated to enhance their writing skills through the course. Such motivation may have contributed to the higher uptake rate of the W&I feedback than those of other systems, as reported in previous studies.

Further studies are needed to investigate how less motivated L2 learners would engage with automated feedback. It should also be acknowledged that the writings produced by our participants in response to the IELTS graph-based prompts were largely predictable, for example, in terms of the vocabulary and sentence structure they would use (Yu et al., 2011), which might be one of the contributing factors for the high accuracy rate of feedback provided by W&I. Further research should use different kinds of writing prompts (e.g., topic-based argumentative essays) which might elicit a wider but less predictable range of lexical and syntactic features, and consequently, there would be a higher challenge on the AWE system to generate targeted, accurate feedback. The statistical analysis showed non-significant differences in participants' engagement with accurate and inaccurate W&I feedback. However, these findings should be interpreted with caution because of (a) the small number of participants in this study and (b) the overall small amount of feedback provided by W&I. Further studies could enlarge their sample size to increase the statistical power when investigating the impact of accuracy and explicitness of different types of feedback on students' engagement with feedback. It is also important to look at the interactive effects of multiple factors (e.g., feedback explicitness and accuracy) on students' feedback engagement (see Guo et al., 2021).

Acknowledgements

We are grateful for the anonymous reviewers' helpful comments and suggestions on earlier drafts of this paper. This research was funded by the TOEFL Small Grants for Doctoral Research in Second or Foreign Language Assessment (2018) from Educational Testing Service, the Assessment Research Awards (2019) from British Council, and DET Dissertation Awards (2020) from Duolingo.

References

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology, 37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal, 100*(1), 320–340.

Bitchener, J. (2017). Why some learners fail to benefit from written corrective feedback. In H. Nassaji, & E. Kartchava (Eds.), *Corrective feedback in second language teaching and learning: Research, theory, applications, implications* (pp. 129–140). Taylor & Francis.

Bitchener, J., & Storch, N. (2016). *Written corrective feedback for L2 development*. Multilingual Matters.

Bolzer, M., Strijbos, J., & Fischer, F. (2015). Inferring mindful cognitive processing of peer-feedback via eye-tracking: Role of feedback-characteristics, fixation-duration and transitions. *Journal of Computer Assisted Learning, 31(5)*, 422–434. https://doi.org/10.1111/jcal.12091

Chapelle, C., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*, 385–405.

Chen, C. F., & Cheng, W. Y. (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. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44145/1/12_02_chencheng.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Erlbaum.

Dikli, S. (2010). The nature of automated essay scoring feedback. *CALICO Journal, 28*(1), 99–134.

El Ebyary, K., & Windeatt, S. (2019). Eye-tracking analysis of EAP participants' regions of interest in computer-based feedback on grammar, usage, mechanics, style and organization and development. *System, 83*, 36–49.

Ellis, R. (2010). A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition, 32*(2), 335–349. https://www.jstor.org/stable/44488131

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80–92.

Gass, S. (1997). *Input, interaction, and the second language learner*. Lawrence Erlbaum Associates.

Guo, Q., Feng, R., & Hua, Y. (2021). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2021.1879161

Hoang, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of My Access. *Language Assessment Quarterly, 13*(4), 359–376.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 1–12. https://doi.org/10.1016/j.asw.2020.100450

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' response to it. *Language Learning & Technology, 19*(2), 50–68. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44417/1/19_02_lavolettepoliokahng.pdf

Limpo, T., Nunes, A., & Coelho, A. (2020). Introduction to the special issue on technology-based writing instruction: A collection of effective tools. *Journal of Writing Research, 12*(1), 1–7. https://doi.org/10.17239/jowr-2020.12.01.01

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *CALICO Journal, 33*(1), 71–91.

Nassaji, H., & Kartchava, E. (Eds.) (2017). *Corrective feedback in second language teaching and learning: Research, theory, applications, implications*. Routledge Taylor & Francis Group.

QSR International Pty Ltd. (2018). *NVivo* (Version 12) [Computer software]. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, 1–16. https://doi.org/10.1016/j.jslw.2021.100816

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. https://doi.org/10.1080/01443410.2015.1136407

Schmidt, R. W. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition, 42*, 1–16.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.

Stevenson, M., & Phakiti, A. (2019). Automated feedback and second language writing. In K. Hyland, & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (pp. 125–142). Cambridge University Press.

Storch, N. (2018). Written corrective feedback from sociocultural theoretical perspectives: A research agenda. *Language Teaching, 51*(2), 262–277.

Tobii Technology. (2017). *Tobii Studio* (Version 3.4.8). https://www.tobiipro.com/

Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback, and teacher feedback in an online ESL writing context. *System, 91*, 1–14.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157–180.

Yu, G., He, L., & Isaacs, T. (2017). The cognitive processes of taking IELTS Academic Writing Task 1: an eye-tracking study. *IELTS Research Reports, 11,* 373–449. https://www.ielts.org/-/media/research-reports/ielts_online_rr_2017-2.ashx

Yu, G., Rea-Dickins, P., & Kiely, R. (2011). The cognitive processes of taking IELTS Academic Writing Task 1. *IELTS Research Reports, 11,* 1–77. https://www.ielts.org/-/media/research-reports/ielts_rr_volume11_report6.ashx

Zhang, Z. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing, 43,* 1–14. https://doi.org/10.1016/j.asw.2019.100439

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36,* 90–102. https://doi.org/10.1016/j.asw.2018.02.004

Appendix A. Writing Prompts

Prompt 1: Book Reading and Types of Book

Take a look at the graphics and complete the task.

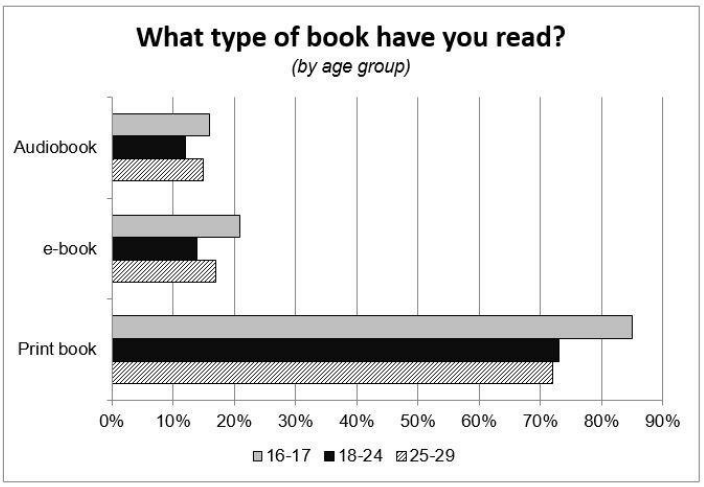
The table and chart show data about reading among people under 30 years old.

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

You should write about 300 words.

Book Reading (2012)

Age Group	Read a Book	Books Read (average)
16-17	90%	12
18-24	79%	11
25-29	81%	11



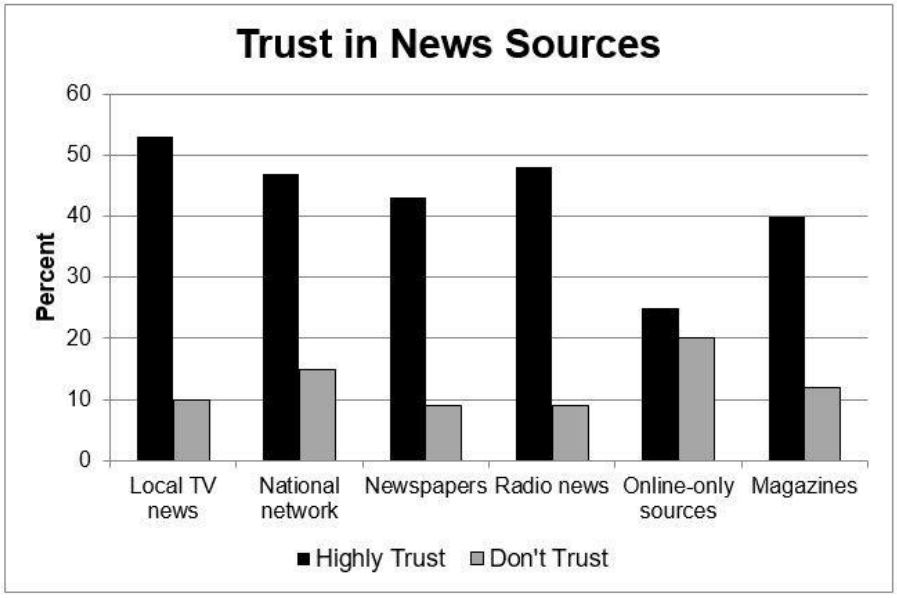
Prompt 2: Trust in News Sources

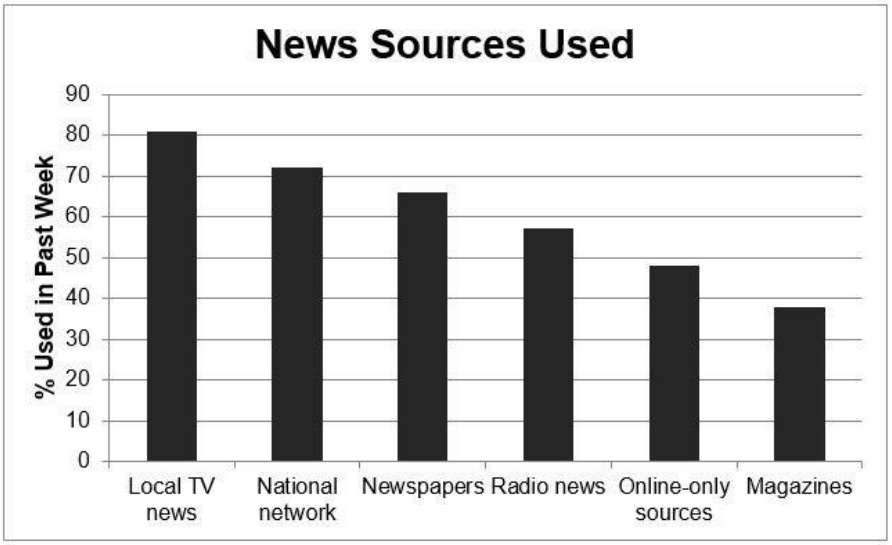
Take a look at the graphics and complete the task.

The charts show where people get their news and how much they trust these sources.

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

You should write about 300 words.




News Sources Used
% Used in Past Week
90
80
70
60
50
40
30
20
10
0
Local TV news
National network
Newspapers
Radio news
Online-only sources
Magazines

Appendix B. Stimulated Recall Interview Guide

1. What did you think when you saw this amount of feedback at the beginning of your essay revision?
2. Why did you fixate on the writing prompt?
3. Why did you fixate on the writing aids?
4. Why did you fixate on the feedback?
5. Why did you dwell on the area of revision?
6. Why did you switch between the area of revision and the writing prompt / writing aids / feedback?
7. Why did you accept / reject / ignore the feedback?
8. Did you understand the underlying rule of the feedback? Why or why not?
9. Why did you make this revision based on the feedback?
10. Why did you not make any revision based on the feedback?
11. Why did you submit your draft?
12. Why did you look away? What happened?
13. You looked happy / excited / confused / annoyed / frustrated / agitated… Why?
14. During the process of essay revision, you asked me "…", why?
15. Do you have any additional comments / questions?

Appendix C. Coding Scheme of Eye Gazeplot Recordings

Codes		Description
Reading Writing Prompt (AOI1)	Reading Instruction Reading Image 1 Reading Image 2	Participant was fixating on the area of the writing prompt
Revising Essay (AOI2)	Pausing Resubmission Revision responses to DWF Revision responses to IWF Revision responses to ISF Self-initiated revisions	Participant was either dwelling in the essay revision area (i.e., pausing), resubmitting their drafts (i.e., Resubmission), or making revisions in response to the automated feedback
Reading Writing Aids (AOI3)	Reading Changes Reading Help Reading History	Participants was fixating on the area of writing aids
Reading Automated Feedback (AOI4)	Reading General Comment Reading Score Reading Your Progress Reading DWF Reading IWF Reading ISF	Participant was fixating on the area of automated feedback
Switching	Between Feedback and Prompt Between Feedback and Revision Between Feedback and Writing Aids Between Prompt and Revision Between Prompt and Writing Aids Between Writing Aids and Revision	Participant switched between the four AOIs
Looking Away		Participant's eyes moved away from the screen or the coder could not infer what the participant was fixating on

Notes. DWF = Direct Word-level Feedback; IWF = Indirect Word-level Feedback; ISF = Indirect Sentence-level Feedback; AOI = Area of Interest. This study did not focus on the General Comment, Score, and Your Progress provided by W&I because such feedback was too general to guide actual revisions.

Appendix D. Coding Scheme of Cognitive Effort Expenditure in W&I Feedback

1. Sample Coding of Cognitive Effort Expenditure

1.1. Direct Word-level Feedback

1.1.1. Participant text: As we can see, people nowadays can get news from a considerably multiple **medias**[1]. (P#6_P2)

1.1.2. W&I feedback: Word ending! The end of this word is unusual. Perhaps "media" is better.

1.1.3. Interview excerpt:

> The Researcher: You clicked open the feedback and read it for a while. Why?
>
> The Participant: Here I was writing about different platforms from which people can get news, so the word "media" should be in its plural form. However, the system flagged it and suggested the use of 'media'. I was a bit confused and trying to recall what should be the correct form.

1.2. Indirect Word-level Feedback

1.2.1. Participant text: **According** to the data, three age groups read a book most of their reading time especially the younger group spent 90 percent of their reading time reading a book. (P#5_P1)

1.2.2. W&I feedback: Suspicious word. Is this the correct word? This doesn't look right to us.

1.2.3. Interview excerpt:

> The researcher: You seemed surprised when seeing the word "According" being flagged?
> The Participant: Yes. I was very surprised. The system said that this word does not look right but it looks right to me. I have used "According to" in my writing all the time and it has never been pointed out by any of my writing teachers. I don't think it is wrong.

1.3. Indirect Sentence-level Feedback

1.3.1. Participant text: However, the percentage of people who have read ebooks and audiobooks is much lower. (P#9_P1)

1.3.2. W&I feedback: The system highlighted the sentence in a crossed lighter-coloured background.

1.3.3. Interview excerpt:

> The researcher: After reading this highlighted sentence, you switched to check Image 2 of the writing prompt. Why?
> The Participant: I was checking whether my summary that the percentage of people who have read ebooks and audiobooks is much lower was accurate. Then I confirmed that I did make an accurate summary, so I decided not to revise the sentence.

[1] The words or phrases flagged by W&I in the student text are highlighted in bold.

2. Sample Coding of No Cognitive Effort Expenditure

2.1. Direct Word-level Feedback

2.1.1. Participant text: In the graph of news sources used, the **percent** goes declined from the local TV news to the magazines. (P#15_P2)

2.1.2. W&I feedback: Perhaps 'percentage' is better.

2.1.3. Interview excerpt:

The researcher: After quickly reading the feedback, you changed all 'percent' to 'percentage'. Do you understand the difference between them?

The Participant: No, I just followed the feedback.

2.2. Indirect Word-level Feedback

2.2.1. Participant text: **On the contrary**, online-only sources has both low percent of trust and use. (P#21_P2)

2.2.2. W&I feedback: Suspicious word. Is this the correct word? This doesn't look right to us.

2.2.3. Interview excerpt:

The researcher: Why did you change "On the contrary" into "However"? Did you think the phrase was wrong?

The participant: I did not know why the phrased was flagged at all. I changed it into "However" to see how the system would respond to it.

2.3. Indirect Sentence-level Feedback

2.3.1. Participant text: The two graphs are about book reading and types of books which are different in three age groups. (P#3_P1)

2.3.2. W&I feedback: The system highlighted the sentence in a crossed lighter-coloured background.

2.3.3. Interview excerpt:

The researcher: I noticed that you deleted the highlighted sentence after scanning it. Why?

The participant: I deleted it because I had no clue what was wrong, and it was annoying to see it highlighted.

Appendix E. Coding Scheme of Revision Type (Adapted from Barkaoui, 2016, p. 326)

(1) Surface Revisions

a. Original draft: Something can be figured out through these statistical **diagram** where relevant. (P#5_P1)
b. Revised draft: Something can be figured out through these statistical **diagrams** where relevant.

(2) Content Revisions

a. Original draft: Additionally, magazines are the most uncommonly used sources compared with other five sources and online-only resources are considered most **reliable**. (P#4_P2)
b. Revised draft: Additionally, magazines are the most uncommonly used sources compared with other five sources and online-only resources are considered most **unreliable**.

Appendix F. Coding Scheme of Revision Quality (Adopted from Bai & Hu, 2017, p. 72)

1. Good Revisions
   a. Original text: Secondly, the average number of books they read per year have little gap among three age groups. (P#8_P1)
   b. Revised text: Secondly, the average number of books they read per year have little gap among **the** three age groups.

2. Neutral Revisions
   a. Original text: **According to** information presented in the two graphs, there are mainly six ways to get news, including local TV news, national network, newspapers, radio news, online-only sources and magazines. (P#23_P2)
   b. Revised text: **Based on** information presented in the two graphs, there are mainly six ways to get news, including local TV news, national network, newspapers, radio news, online-only sources and magazines.

3. Bad Revisions
   a. Original text: **According to** the graph information included in the first bar chart, we can discern how much people trust or not trust about the news sources while in the second bar chart we can have a more specifically data of people's usage rate from each news source in the past week. (P#12_P2)
   b. Revision draft: **Accord to** the graph information included in the first bar chart, we can discern how much people trust or not trust about the news sources while in the second bar chart we can have a more specifically data of people's usage rate from each news source in the past week.

About the Authors

Sha Liu is a PhD candidate at the University of Bristol. Her work has been published in journals including the *Computer Assisted Language Instruction Consortium (CALICO)* and *Assessing Writing*. Her research interests include second language writing assessment, automated writing evaluation, eye-tracking, and Rasch modelling. See https://orcid.org/0000-0002-0439-660X for a list of her publications.

E-mail: sha.liu@bristol.ac.uk

Guoxing Yu is a Professor of Language Assessment at the University of Bristol. He is an Expert Member of the European Association for Language Testing and Assessment. He has directed several research projects on IELTS and TOEFL iBT. See https://orcid.org/0000-0001-5997-2734 for a list of his publications.

E-mail: guoxing.yu@bristol.ac.uk