The efficacy of automated feedback on summary content in an online formative assessment module for undergraduate L2 writing instruction in Japan

Yasuyo Sawaki
Waseda University, Japan

Yutaka Ishii
Chiba University, Japan

Yoko Oi
Seisen University, Japan

The present mixed-methods study examined the efficacy of W-Writing, an online formative assessment module for summary writing designed for supplemental use in undergraduate introductory academic writing courses in Japan to enhance summary writing instruction under the new normal that emerged after the COVID pandemic. This module automatically delivered four types of content feedback inspired by the formative assessment and learning-oriented assessment (LOA) literature during the revision stage of a summary writing task. The efficacy of the immediate content feedback was examined from two perspectives: students' revision behaviors and students' and instructors' perceptions of the W-Writing module. Students enrolled in the course and four course instructors participated in this study. A frequency analysis of revision operations applied to the initial summary draft of 60 randomly selected students during revisions suggested that they were engaged in meaningful revision activities. The students' rating of the usefulness of W-Writing (n = 81) was also high. These results were corroborated by results from a thematic analysis conducted on instructors' interview data, which suggested the instructors' high endorsement of W-Writing's usefulness and their perceptions

Email address for correspondence: ysawaki@waseda.jp

of students' high engagement in W-Writing activities. The above results provide initial empirical support for the efficacy of the W-Writing feedback design.

Keywords: automated writing evaluation (AWE), formative assessment, learning-oriented assessment (LOA), technology-mediated feedback, summary writing

Introduction

The COVID pandemic required many higher education institutions worldwide to make a sudden shift from in-person instruction to online instruction. The adoption of technology for instruction during the pandemic often aimed to avoid disruption of learning in *ad hoc* remote instructional environments. Under the new normal which emerged after the protracted process of recovery from the pandemic, however, technology is employed by choice, often with the aim of optimizing learning through hybrid instruction that flexibly combines online and in-person environments.

While various technological capabilities can be leveraged for instruction, a particularly attractive one is technology-mediated feedback, whose adoption to support second language (L2) learning has accelerated in recent years. A notable strength of technology-mediated feedback is the tight control of its timing and content (Jang & Sinclair, 2022; Keppell & Carless, 2006; Voss, 2021). For instance, while it is time-consuming for an instructor to annotate a large number of learners' essays, automated text analysis can offer feedback on specific aspects of learners' performance instantly for delivery at a pre-determined time during a learning program. However, the efficacy of technology-mediated feedback for formative purposes requires a focused empirical investigation. In particular, the degree to which technology-mediated feedback gets acted upon and is perceived as a useful aid that supports learning by students and their course instructors is key to justifying its use in the L2 classroom (Bachman & Damböck, 2018).

The present mixed-methods study examined the efficacy of technology-mediated feedback provided by W-Writing, a web-based module for formative assessment of summary writing skills developed for supplemental use in introductory undergraduate English academic writing courses in Japan. The conceptualization of W-Writing started as part of a larger research project in 2019, before the pandemic, while its development was facilitated by the rapid increase in demand for online instructional materials during the pandemic and in the new normal. Despite the fact that fostering basic source-based writing skills is a major learning goal for many introductory L2 academic writing programs, achieving it is often a challenge. Primary reasons include the lack of time for instruction and practice, as well as the complexity of the summarization construct, which makes the provision of quality feedback a daunting task for instructors. Accordingly, W-Writing was designed to offer focused summary writing practice opportunities with immediate feedback on key aspects of summary writing practice to assist both learners and course instructors. After repeated piloting and refinement of the module during the pandemic, this assessment module inspired by the formative assessment and learning-oriented assessment (LOA) literature was introduced to a hybrid L2 academic writing course at a university in Japan in spring 2023. The present paper reports on the results of an initial validation study conducted in the course, focusing specifically on students' revision behaviors upon provision of the automated feedback and course instructors' and students' perceptions of their experience of using the module as part of the L2 writing course.

Literature review

Formative assessment is an ongoing process where assessment results are utilized by teachers and learners to make various types of decisions to adjust instruction and learning, helping learners achieve their learning goals and enhance their knowledge, skills, and abilities (Black & Wiliam, 1998; Turner & Purpura, 2016). Feedback presented to learners for formative purposes must be carefully designed, given the pivotal role feedback plays in the learning process. In the context of LOA, Turner and Purpura (2016) specified feedback as the central component of the learning dimension of their seven-dimensional working model of learning-oriented language assessment. This reflects the critical importance of the learner's understanding of their current level, the target level, and how to close the gap between the two to facilitate learning (Hattie & Timperlay, 2007; Sadler, 1989). Previous studies have proposed various features of effective feedback (Boud & Molloy, 2013; Carless, 2007; Keppell & Carless, 2006; Lam, 2021). For instance, according to the principles of effective feedback practice summarized by Lam (2021), it is specific, forward-looking, and appropriate in level; it also relates to success criteria, engages the learner in a proactive role, and promotes self-regulated learning.

When technology is introduced to a formative assessment process, an expectation would be that learners and instructors benefit from real-time updates on learning progression throughout a course, enabling immediate feedback provision and remedial actions, even in a remote instructional environment (Choi & McClenen, 2020). Indeed, an oft-mentioned premise of introducing technology-mediated feedback for automated writing evaluation (AWE) is the timeliness of feedback. Making automated feedback available to learners when they require it would not only encourage them to use the information for further learning (Boud & Molloy, 2013; Carless & Boud, 2018; Keppell & Carless, 2006; Lam, 2021; Voss, 2021) but also contribute to building a sustainable feedback practice that alleviates the resource-intensiveness of instructors' work required for feedback provision (Boud & Molloy, 2013). Consistent with this view, empirical studies concerning AWE have been fast accumulating in applied linguistics in recent years. Reflecting the tradition of written corrective feedback research in the field, previous AWE studies in L2 writing have focused primarily on grammar and vocabulary (Biber et al., 2011; Fu et al., 2022; Stevenson & Phakiti, 2014), with a bulk of empirical studies examining the efficacy of corrective feedback as implemented in commercial AWE tools such as Criterion®, Grammarly, and Pigai. The rapid expansion of studies on AWE has been documented in recent research syntheses (Barrot, 2024; Ding & Zou, 2024; Fan & Ma, 2022; Fu et al., 2022; Ngo et al., 2024; Shi & Aryadoust, 2024; Stevenson & Phakiti, 2014). These works have pointed out some drawbacks of previous AWE tools, such as overcorrection of linguistic errors and the superficial nature of errors that can be identified. Yet, they have commonly reported generally positive effects of AWE tools on various aspects of L2 writing performance, including writing accuracy and writing quality, particularly when they are used for a certain duration, such as for one semester (Ding & Zou, 2024).

At the same time, however, the efficacy of AWE is underexplored for other outcome measures. One example is learner feedback engagement (Barrot, 2024), yet this topic has drawn the attention of the field in recent years. An influential framework recently proposed by Zhang and Hyland (2018, 2023) conceptualizes feedback engagement as a multifaceted construct comprising behavioral, cognitive, and affective components.

Emerging qualitative case studies (Koltovskaia, 2020; Zhang & Hyland, 2018) based on this framework have offered valuable insights into the importance of learner perspectives in evaluating AWE effectiveness. Zhang and Hyland's (2018) study of two Chinese learners demonstrated a key mediating role that engagement plays in learner feedback use and its impact on writing progress. For instance, when AWE feedback was compared against teacher feedback, the former led to greater flexibility in the pace of revisions due to the increased autonomy felt by the learner. Meanwhile, Koltovskaia (2020) explored the engagement of two English-as-a-second-language (ESL) college students in the corrective feedback provided in Grammarly. It was found that the two participants' levels of cognitive engagement differed, one exhibiting a higher level by attempting to address challenging feedback and the other exhibiting a lower level suggested by uncritical feedback acceptance. Indeed, descriptive results based on multiple data sources obtained in these studies are of great value, but more studies involving larger samples and in L2 classrooms are required to gauge the generalizability of the available study findings (Cerezo et al., 2019; Park & Kim, 2019).

Another area with limited empirical evidence for the efficacy of technology-mediated feedback is learners' and instructors' perceptions of AWE feedback. On the one hand, recent systematic reviews by Ding and Zou (2024), Shi and Aryadoust (2024), and Wilson et al. (2024) showed that empirical results on learners' perceptions so far are mixed, which may also be explained by various learner and instructor variables. On the other hand, results of instructor perception studies are generally positive (Shi & Aryadoust, 2024). However, given the small number of available studies, this area has been identified as a major gap in the literature (Ding & Zou, 2024; Koltovskaia, 2023). This is a concern because how these key stakeholder groups view such feedback is likely to affect the degree to which they utilize such feedback and how they incorporate it into their instruction and learning.

To sum up, the review of the studies above suggests that AWE feedback efficacy for certain aspects of L2 writing performance has obtained some empirical support, while further explorations are required on the effects of AWE feedback on other outcome measures, such as L2 learner feedback engagement and learners' and instructors' perceptions of AWE feedback. While the positive results on the efficacy of AWE are encouraging, there are two specific issues that require further scrutiny from the perspective of the present study. Firstly, it is worth noting that the bulk of the literature has dealt with corrective feedback on vocabulary and grammar, whereas feedback on content, particularly that concerning source-based writing, has not drawn much research attention. Previous systematic reviews (Fan & Ma, 2022; Fu et al., 2022; Stevenson & Phakiti, 2014) have identified only a few primary studies on content feedback, all of which concerned the feedback available in an AWE tool called *Summary Street*, now a component of Pearson's WriteToLearn®. This program delivers a series of content feedback including measures of semantic similarity between a given learner-produced summary and the source text based on latent semantic analysis (Landauer & Dumais, 1997). Classroom trials conducted in US grade schools (Foltz et al., 2013; Wade-Stein & Kintsch, 2004) reported positive effects of this tool on learners' summary performance and task engagement, as measured by time on task (time spent by a learner on a given task). Yet, with this small number of previous investigations and the lack of L2 writer performance data, it is fair to say that empirical evidence on the efficacy of content feedback is severely limited in the AWE literature.

Secondly, discussion on design features of AWE tools that facilitate learner feedback engagement is rather scarce in the literature. Generally speaking, the currently available AWE tools focus primarily on the delivery of fine-grained individualized feedback per se, relying heavily on instructors to design teaching and learning activities to promote learner uptake. This may not be an issue when an AWE tool is used in an in-person environment. However, when designing AWE tools that can accommodate various types of in-person and online learning environments for the new normal, it might be worth considering a more structured tool design that incorporates activities that promote learner feedback engagement even with limited instructor assistance—specifically, following the current view of the learner as an active agent who not only receives feedback but also interprets it and utilizes it for learning (Andrade, 2010; Black & Wiliam, 1998; Boud & Molloy, 2013; Carless & Boud, 2018; Hattie & Timperley, 2007; Lam, 2021). Hence, a feedback design that promotes active learner engagement is in order. As a way forward, two principles raised in the formative assessment and LOA literature may inform the design of such tools. One is to conceptualize feedback practice as a curriculum issue, where careful planning is required to determine how feedback should be embedded in a larger cycle(s) of learning tasks to create an environment that facilitates learner uptake (Boud & Molloy, 2013; Carless, 2007). One way to do this would be to devise a module comprising multiple parallel writing tasks, where the learner is encouraged to utilize automated feedback presented during the revision stage of a given task to do better on subsequent tasks. Embedding the feedback in a learning process in such a way would encourage the learner to attend to and act upon the feedback provided.

Another principle that would inform AWE tool design is to devise tasks where the learner plays the role of an assessment agent to foster co-regulated learning (Andrade, 2010; Andrade et al., 2021). Built on the concept of self-regulated learning, co-regulated learning is defined by Andrade et al. (2021) as "the interactive, reciprocal influences of all sources of information about learning [such as the learner, peers, instructors, and technology] on each other" (p. 2). Indeed, certain AWE tools today such as Criterion and Pigai involve peer assessment functions as supplements, whereas discussions on the role of peer assessment (Ding & Zou, 2024) as well as self-assessment in AWE appear scant. Accordingly, it may be worth considering the adoption of functions for peer and self-assessment as an integral part of AWE tools. Such activities, which assign the learner an agentive role, offer valuable sources of information that would help the learner reflect on their current level, the target level, and how to close the gap between the two (Hattie & Timperley, 2007; Sadler, 1989) as they move through the phases of co-regulated learning (Andrade et al., 2021).

With the current state of the AWE literature reviewed above as the background, the present study examined the efficacy of automated content feedback for summary writing instruction delivered through W-Writing, an online module designed for L2 academic writing course use at universities in Japan. W-Writing feedback was designed specifically with the two above-mentioned design principles informed by the formative assessment and LOA literature in mind in order to promote learners' feedback engagement. The present study was one of the two initial validation studies of the W-Writing module. While the other study (Sawaki et al., 2024) examined changes in learner written summary products across occasions of W-Writing use quantitatively, the present mixed-methods study explored the W-Writing feedback quality from two perspectives: learners' revision behaviors as they received W-Writing feedback as scaffolding (Zhang & Hyland, 2018, 2023); and how this behavioral aspect of learners' feedback engagement may be explained by learners' and instructors' perceptions of the W-Writing module. The following two research questions were addressed:

1. How and to what extent do students revise their summaries based on the W-Writing feedback?
2. How do learners and instructors perceive their experience of using the online module as part of summary writing instruction and learning?

Method

Overall study design

The present study adopted a convergent mixed-methods study design (Creswell, 2021), where data to address the two key research questions were collected concurrently. On the one hand, quantitative data were collected from students enrolled in the target course to address both research questions: students' revision behaviors based on the W-Writing summary content feedback (Research Question 1) and their perceptions of the usefulness of the feedback (Research Question 2). On the other hand, qualitative data obtained through semi-structured interviews with course instructors served as the main data source to explore instructors' perceptions of the usefulness of W-Writing feedback (Research Question 2). Furthermore, the instructor interview data offered supplementary information facilitating interpretation of the quantitative data collected from students to address both research questions.

Study context

The target course in which data for this study were collected was an introductory academic writing course required for first-year undergraduate English majors in the school of education at a private university (University A) in Tokyo. This one-semester course, which met for 100 minutes per week for 14 weeks, focused on paragraph writing, summary writing, and an introduction to essay writing. Two weeks during the semester were allocated to summary writing. W-Writing was integrated into the regular course syllabus to provide learners with focused summary writing practice opportunities with immediate feedback to enhance learning of this basic source use skill. All summary writing tasks completed by students were treated as writing assignments for the course. Given the formative nature of the W-Writing activities, the summary writing tasks carried no or small weight in the course grade, depending on the grading policy of the course instructor. At the time of study participation, which took place in Weeks 8 and 9, all classes had completed all planned textbook units on paragraph writing in different rhetorical structures (e.g., classification, cause/effect, comparison/contrast). The classes had no previous exposure to summary writing instruction in the course.

Participants

Study participants were instructors and their students enrolled in seven classes (two advanced and five intermediate) of the target course in spring 2023. In each class, the course instructor made an announcement regarding this study and asked students' permission for research use of data that they generated through their regular class activities. The students were permitted to opt out of the study if they wished not to participate. The students provided their responses concerning their study participation online, which were not disclosed to the course instructors. Consent for study participation was obtained from 101 of the 126 students enrolled. Most of them were Japanese L1 first-year undergraduate students educated in Japan, while a few international students and returnees were also included. Information about students' gender and age was not available. The English proficiency level of the student population was primarily at the B1 level on the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001).

As for course instructors, the first author contacted all of them via e-mail to solicit their participation in the semi-structured interviews. Originally, five instructors (an advanced-level section instructor and four intermediate-level section instructors) agreed to participate. All were Japanese L1 speakers with functional English language proficiency and with doctorates in applied linguistics and related fields. They were all experienced instructors of academic writing at the undergraduate level in Japan. However, interview data obtained from an intermediate-level section instructor was unavailable for detailed analyses due to recording error. For this reason, interview results only for the remaining four instructors will be reported subsequently.

Instruments

The materials described below were involved in class activities associated with the W-Writing module. These materials were developed as part of the larger project and piloted in academic writing courses at another private university (University B) in Tokyo in 2021 and 2022 (All appendices noted below are available from the Open Science Framework website: https://osf.io/pv4e8/?view_only=ff5be79463124a649efd60142a3ebcb7; See also Sawaki et al., 2024 for further details on these materials and their theoretical background).

W-Writing

This web-based module for formative assessment of summary writing was programmed to administer summary writing tasks with automated feedback on the content of learner-produced summary responses as scaffolding for revisions. The content feedback implemented in W-Writing focuses on the macrorules of summarization, or "semantic mapping rules" (Kintsch & van Dijk, 1978, p. 366) that the reader is hypothesized to deploy to generate a text gist, proposed in the construction-integration model of text processing (Kintsch, 1998; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). As the basis of the W-Writing feedback design, the macrorules were defined in terms of the five rules of summarization below, following Hare and Borchardt (1984):

- Rule 1: Collapse lists.
- Rule 2: Omit unnecessary details.
- Rule 3: Use topic sentences.
- Rule 4: Collapse paragraphs.
- Rule 5: Polish the summary (use connecting words and paraphrasing).

The user of W-Writing can customize the module content to be delivered to the learner by incorporating source texts of their choice for the summary tasks and modifying the feedback reflective of the specific source text content. Further details regarding the content feedback delivered through W-Writing in the present study will be provided in the Procedure section.

Summarization strategies instruction video and handout

A 25-minute video delivered in Japanese by the first author was prepared for instruction on summarization strategies. In the video, a definition of summarization was presented and the importance of summarization skills in academic work explained. After that, the five rules of summarization described above were presented, along with concrete examples. A handout noting the definition of the five rules and corresponding concrete examples presented in the video lecture was also prepared for students' reference.

Source texts for the summary tasks

Two source texts (Appendices A and B; the same texts as those employed by Sawaki et al., 2024) were devised for use as part of the summary tasks delivered on W-Writing. Text A was a problem-solution text on a new environmental technology for generating drinking water, while Text B was on recent research on the effects of napping on learning. Both texts were comparable in readability and length (Flesch-Kincaid Grade Level of 10.5 for Text A, and 9.3 for Text B; 391 words for Text A, and 415 words for Text B).

Learner questionnaire

This questionnaire comprised items on the usability of the instructional materials and the usefulness of the W-Writing module. Among the 18 items in total, results on only one section concerning the usefulness of summarization instruction and feedback for learning were analyzed in this study. The seven items were statements written in Japanese on a five-point Likert scale (1 = strongly disagree; 5 = strongly agree). Four items were on the usefulness of the four types of content feedback delivered through W-Writing (e.g., "The function to compare my text highlights and those of an expert summarizer was useful."), while the other three were statements on the usefulness of W-Writing for understanding what summarization is, for understanding the five rules of summarization, and for future learning (e.g., "What I learned in this module is helpful for my future learning.").

Procedure

The summary writing activities took place in Weeks 8 and 9 of the target course. Five of the seven classes (one advanced and four intermediate) met on Zoom, while the remaining two classes (one advanced and one intermediate) met in person in both weeks. Table 1 summarizes the procedure each class followed. The instructor provided directions to students to complete the activities scheduled for the day. The instructor was present throughout the W-Writing activities in class to support the students as needed. Two teaching assistants were also on standby to handle technical issues. It should be noted that, while all classes conducted all activities in Table 1 in the same order, the exact schedule differed across classes. In some classes, students completed Activity 1 as an assignment, while the entire class time in Week 8 was allocated to Activities 2 and 3, and the first half of Week 9 class time to Activities 4 and 5. In the others, the entire class time was allocated to Activities 1 and 2 in Week 8, and to Activities 3, 4, and 5 in Week 9. This means that the intervals between Activities 2 and 3 and between Activities 3 and 4 were not strictly controlled. Despite this, the total time spent covering all parts was roughly the same across classes (approximately 200 minutes, including time required for giving necessary directions).

Table 1. Summary writing activity procedure (Weeks 8 and 9)

Step	Activity
Activity 1 (40 min.)	Summarization instruction • Learning five rules of summarization • Explanation of the summarization rubrics and goal setting
Activity 2 (40 min.)	Learning task (Summary Task 1) • Write a summary on your own (without scaffolding)
Activity 3 (40-50 min.)	Learning task (Summary Task 1) • Revise your summary (with scaffolding)
Activity 4 (40 min.)	Transfer task (Summary Task 2) • Check your learning (no scaffolding)
Activity 5 (10 min.)	Learner questionnaire survey

The specific activities conducted in each class were as follows. Firstly, as Activity 1, students watched the summarization instruction video alongside the handout, which were delivered through Dropbox or the learning management system (LMS). After the instructor provided additional explanations and responded to students' questions as needed, students logged onto W-Writing to complete Activities 2 through 4. The order in which the two source texts were presented in Summary Task 1 (Activities 2 and 3) and Summary Task 4 (Activity 4) was randomized across classes. This counter-balancing was necessary in order to take the passage effect into consideration in comparing the initial and final drafts of students' summaries to address Research Question 1. As Activity 2, the students completed Summary Task 1, where they were instructed to read the source text and write a summary of it in approximately 80 words. Dictionary use was allowed. As this task was designated as the learning task, the students completed it without scaffolding first. Once the response was submitted to the system, the student moved on to Activity 3, where they revised the initial draft for Summary Task 1 while reviewing four different types of content feedback below as scaffolding. In W-Writing, all four different types of content feedback were presented automatically to all learners in a fixed order (for screenshots of the W-Writing feedback, see Appendices C through F; see also Sawaki et al., 2024):

- *Feedback 1 (Check source text comprehension)*. W-Writing prompted the student to review main ideas of the source text by highlighting important parts of the text that need to be included in a summary using the highlighting function on the screen. When the student clicked the "Compare" button, their highlights were displayed along with underlines of text parts selected by the Expert Summarizer for their importance. (These underlines were based on the 25 percent most important idea units of the source text identified independently by six raters and topic notes developed by raters in the larger study.)

- *Feedback 2 (Check summary outline)*. The student was prompted to enter a text outline of their summary in a text box on the screen. When the student pressed the "Compare" button, an outline prepared by the Expert Summarizer was presented side by side with the student's own. (The Expert Summarizer outline was a short version of the topic notes used by raters for scoring.)

- *Feedback 3 (Self-assess summarization rule use)*. A self-assessment checklist was presented on the screen, where the student was instructed to review each item on the list to self-assess the appropriateness of the summary content. This checklist comprised 14 items targeting the use of the five rules of summarization presented in the video lecture.

- *Feedback 4 (Check paraphrasing)*. This feedback employed an automated analysis of the learner-produced summary, flagging text strings of three words or longer identified as verbatim copies from the source text. The parts identified in the student summary and the corresponding parts in the source text were highlighted in red on the screen. The student was instructed to review the highlighted parts and modify them as appropriate.

After submitting the final draft of Summary Task 1 at the end of Activity 3, the student proceeded to Summary Task 2 (Activity 4). Parallel in design to Summary Task 1, Summary Task 2 was designated a transfer task, which the student completed without scaffolding to check how well they could apply the instructional content (the five rules of summarization) to a new task. Finally, the student completed the student questionnaire through the link to Google Form shared by the instructor (Activity 5).

Instructor interviews

Semi-structured interviews were conducted in August and September 2023 with the four instructors who agreed to participate in this study. All interviews were conducted on Zoom to accommodate the instructors' schedules during the summer break. While this was more than two months after the summary writing classes took place during the semester, this timing (after the grading period) was chosen to ensure that the instructors could reflect on the role of W-Writing-related summary writing activities in their academic writing instruction and assessment during the semester. The interview questions (Appendix G) mainly concerned the instructors' perceptions of (1) the appropriateness of the procedure and schedule for conducting class activities based on W-Writing; (2) student engagement in W-Writing activities; and (3) the appropriateness, usefulness, and issues of W-Writing for teaching and learning. For eliciting the instructors' comments on the last of these three topics, the summarization strategies instruction handout and each screen of the W-Writing interface were presented to them, one by one. The first author served as the interviewer. All interviews were conducted in Japanese, the first language of the interviewer and the interviewees. Each interview lasted approximately one hour. The interviews were recorded and transcribed using the Zoom transcriber function.

Analyses

Multiple quantitative and qualitative data sources were triangulated to address each research question.

Analyses concerning Research Question 1 (Learners' summary draft revision behaviors based on the W-Writing content feedback)

Two types of quantitative analyses were conducted to compare the initial and final drafts of Summary Task 1 submitted at the end of Activities 2 and 3, respectively, by 60 students (30 each for Forms 1 and 2) selected randomly across the seven sections. Firstly, using Microsoft Word's Compare function, the initial and final drafts were combined to generate one document for each student with all differences between the two drafts marked. The resulting document was segmented into orthographic sentences and analyzed for the observed revision operations applied to the initial draft (to generate the final draft) using the coding scheme shown in Table 2, which appears in the Results section. This coding scheme comprised three major categories. Category A, adapted from Link et al. (2020), concerned the type of revision act performed (e.g., deletion, addition, change). Categories B and C, developed for the purpose of this study, captured the nature of the revision. Assigning multiple codes to a given sentence was allowed within and across these two categories. Category B comprised surface changes—namely, revisions of vocabulary, grammar, usage, and mechanics. Category C captured changes to the summary content, relevant to the immediate content feedback provided during Activity 3. Substantive revisions directly related to Feedback 1 (*Check source text comprehension*), Feedback 2 (*Check source text outline*), and Feedback 3 (*Self-assess summarization rule use*) were coded. As described below, Feedback 4 (*Check paraphrasing*) required a separate analysis.

Three coders were involved in the revision analysis coding. All students' revisions from the initial to the final summary drafts were coded for Categories A and B by the first and second authors, and for Category C by the first author and another coder. All three coders were Japanese L1 speakers with doctoral degrees in applied linguistics who had extensive academic writing teaching experience. Of the 60 students' revision data, a randomly-selected ten were double coded for initial coding practice. Extensive discussions were held to clarify the specific coding rules deployed and to resolve coding discrepancies. This was followed by independent double coding of another 20 students' data, where acceptable levels of rater agreement were achieved (81.0 percent for Category A, 86.6 percent for Category B, and 80.2 percent for Category C). All coding discrepancies were resolved by discussion. After that, the second author coded the remaining 30 students' revisions for Categories A and B, and the first author for Category C.

As for Feedback 4 (*Check paraphrasing*), an automated analysis was conducted separately on the initial and final drafts to identify verbatim copies (exact matches) of three-word text strings or longer. Then, descriptive statistics (number of exact matches, total words in exact matches, and mean exact match length) were compared between the initial and final drafts to examine the degree to which copied strings had been modified.

Analyses concerning Research Question 2 (learners' and instructors' perceptions of their W-Writing experience)

Firstly, descriptive statistics were calculated for questionnaire responses to the seven items on the usefulness of the W-Writing module, to which only 81 of the 101 student participants responded. Note that the survey response data for the 60 students for whom the revision analyses above were conducted could not be isolated because the survey was anonymous. Cronbach's alpha obtained across the seven items was .794, suggesting satisfactory internal consistency reliability.

Secondly, thematic analysis (Braun & Clarke, 2006) was conducted on the interview protocols obtained from the four instructors. Transcripts obtained from Zoom were manually edited to correct speech recognition errors. The third author then divided the interview transcripts into segments. A segment represented a single episode, typically comprising consecutive sentences within and across turns. Then, following Braun and Clarke's (2006) steps of thematic analysis, the third author coded all segments. The general approach was data-driven in that codes, along with themes and subthemes, were obtained inductively. The coding results were reviewed by the first author, followed by discussions between the first and third authors to resolve discrepancies in the interpretation of the segments. All participants reviewed an earlier draft of the present paper and confirmed accuracy of the interpretation of the results.

Results

Research Question 1: How and to what extent do students revise their summaries based on the W-Writing feedback?

Revision acts

Table 2 summarizes revision acts made by the 60 randomly-selected students during Activity 3. In total, 390 sentences (197 for Text A, 193 for Text B) were analyzed. Due to the similarity of the patterns observed between the texts, only the results for all 60 students will be described below. As can be seen in Table 3, at least one change was made to 279 of the 390 sentences (71.5%) regardless of the revision content. A majority of the revision operations (Categories A1–A5) were changes made to existing information by replacement (observed in 140 sentences in total), as well as deletions and additions of words, phrases, and clauses within sentences (107 and 105 sentences, respectively). In contrast, more major changes such as deleting and adding entire sentences were relatively few (23 and 43 sentences, respectively).

Table 2. Revision acts: coding results

Code	Category	Definition	Text A	Text B	Total
Summary					
	No change	No change was made to the sentence.	63	48	111
	Change	At least one change was made to the sentence.	134	145	279
	Total		197	193	390
Revision acts					
A1	Delete-sentence	The student removed an entire sentence.	7	16	23
A2	Delete-within sentence	The student deleted a word(s), a phrase(s), or a clause(s).	60	47	107
A3	Add-sentence	The student added an entire sentence.	15	28	43
A4	Add-within sentence	The student added a word(s), a phrase(s), or a clause(s)	57	48	105
A5	Change	The student replaced a word(s), a phrase(s), or a clause(s) with an alternative(s)	70	70	140

Note. Number in each cell = frequency of sentences coded.

Revision content

Table 3 summarizes revision content coding results—that is, the nature of revisions made to the 279 sentences involving at least one change. As shown in the top section, 84 sentences involved only surface changes, whereas 195 sentences included at least one content change appearing with or without surface changes. Most surface changes (Categories B1–B7) were accounted for by two types: grammar changes (B5; e.g., correcting grammatical errors, making minor changes to grammar) and word choice changes (B7), which were applied to 39.8 percent and 43.0 percent of the 279 sentences in summaries of Texts A and B, respectively.

Meanwhile, more variation was observed in the type of content change (Categories C1–C12), which is relevant to the W-Writing content feedback. The most frequently observed changes were omitting unnecessary details (C5; 75 sentences) and adding important details (C6; 80 sentences). Moreover, topic sentences existing in the source texts were successfully incorporated into the revised summaries in some cases (C3; 11 sentences). All these operations lead to earning credits in scoring. However, such attempts were not always successful, as can be seen in the occurrences of adding unnecessary details (C7; 31 sentences) and omitting important pieces of information (C8; 17 sentences). While the sentences coded for these four categories were typically paraphrased sufficiently, there were cases where the revisions involved text strings that were highly similar to the language that appeared in the W-Writing feedback. The cases in point are 22 sentences coded for C1, which involved verbatim or near verbatim adoption of the Expert Summarizer text underlines presented in Feedback 1, and seven sentences coded for C2, which employed strings of text displayed only in the expert outline presented as Feedback 2. It is worth noting that major content changes such as creating a topic sentence for a source-text paragraph not including one (C4) and writing a sentence to combine information across paragraphs (C11) were virtually nil. Finally, omitting or synthesizing examples (C9) and omitting a list (C10) were non-existent, partly because few initial drafts included these pieces of information in the first place.

Table 3. Revision content: Coding results

Code	Category	Text A	Text B	Total
Summary				
	The sentence includes a surface change(s) only.	40	44	84
	The sentence includes a content change(s) only.	32	51	83
	The sentence includes both a content change(s) and a surface change(s).	62	50	112
	Total	134	145	279
Surface changes				
B1	Spelling	10	9	19
B2	Capitalization	2	0	2
B3	Punctuation	0	0	0
B4	Sentence structure	2	1	3
B5	Grammar	52	59	111
B6	Formatting	1	1	2
B7	Word choice	67	53	120
Content changes				
C1	Feedback 1: Expert Summarizer text underlines	7	15	22
C2	Feedback 2: Expert Summarizer outline	4	3	7
C3	Feedback 3: Topic sentence use	5	6	11
C4	Feedback 3: Topic sentence creation	0	0	0
C5	Feedback 3: Omit unnecessary details	36	39	75
C6	Feedback 3: Add important details	40	40	80
C7	Feedback 3: Add unnecessary details	15	16	31
C8	Feedback 3: Omit important details	7	10	17
C9	Feedback 3: Omit or synthesize examples	0	0	0
C10	Feedback 3: Omit list	0	0	0
C11	Feedback 3: Combine paragraphs	1	0	1
C12	Feedback 3: Overall organization	0	1	1

Note. Number in each cell = frequency of sentences coded.

Paraphrasing

Table 4 summarizes descriptive statistics for text strings identified as exact copies from the source texts corresponding to Feedback 4. As can be seen in the table, the grand mean for the three measures (number of exact matches, total words in text matches, and mean exact match length) reduced from the initial draft to the final draft. This is consistent with the noticeable decrease of group means for all three measures in the revised draft for Text A (from M = 3.97 to M = 2.77, from M = 17.93 to M = 11.60, and from M = 3.54 to M = 2.97, respectively). In contrast, the results for Text B differed. The group means were even slightly higher for the final version for number of exact matches (M = 3.73 in the initial draft, M = 3.90 in the final draft) and total words in text matches (M = 18.63 in the initial draft, M = 19.73 in the final draft), although the group mean of mean exact match length reduced slightly (M = 3.72 in the initial draft, M = 3.41 in the final draft).

Table 4. Descriptive statistics for exact copying (Feedback 4)

Measure		Text (*n*)	Summary Task 1 *M (SD)* Initial draft (Activity 2)	Summary Task 1 *M (SD)* Final draft (Activity 3)
Exact copying	Number of exact matches	1 (30)	3.97 (2.23)	2.77 (2.01)
		2 (30)	3.73 (2.34)	3.90 (2.96)
		Total (60)	3.85 (2.29)	3.33 (2.59)
	Total words in exact matches	1 (30)	17.93 (13.58)	11.60 (9.73)
		2 (30)	18.63 (15.49)	19.37 (17.20)
		Total (60)	18.28 (14.57)	15.48 (14.50)
	Mean exact match length	1 (30)	3.54 (1.27)	2.97 (1.55)
		2 (30)	3.72 (1.51)	3.41 (1.95)
		Total (60)	3.63 (1.40)	3.19 (1.77)

Research question 2: How do learners and instructors perceive their experience of using the online module as part of summary writing instruction and learning?

Student survey results

Figure 1 summarizes students' ratings on the seven survey items regarding the usefulness of the four types of immediate feedback and their experience of using the W-Writing module for learning summary writing (n = 81). As can be seen in the figure, the students' ratings were high across the board, the percentages of students endorsing the usefulness (the percentages of rating 4 [agree] and rating 5 [strongly agree] combined) being in the upper 80s or above. Concerning the four types of immediate feedback, the endorsement of Feedback 1 (text underline, 92.6%) and Feedback 2 (outline, 92.6%) was higher than that of Feedback 3 (checklist, 86.4%) and Feedback 4 (exact match, 87.7%). In addition, over 90 percent of the students agreed upon the usefulness of W-Writing for understanding what summarization is (90.1%) and the five rules of summarization (92.6%), as well as for their future learning (93.8%).

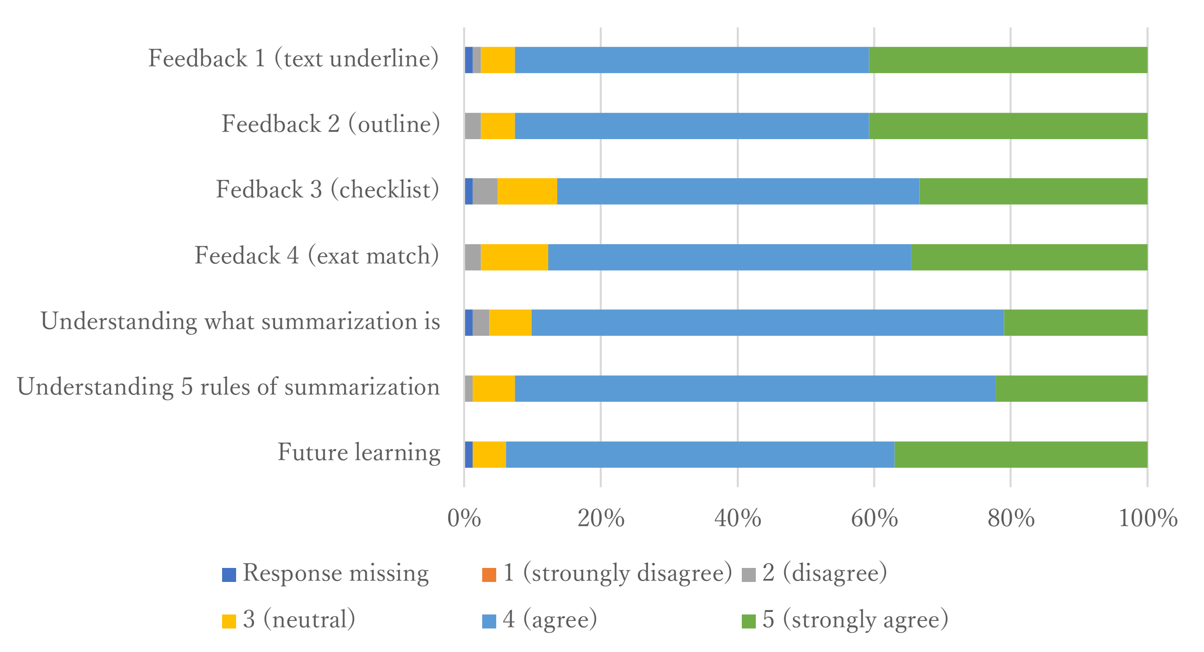

Figure 1. Learner agreement ratings on the usefulness of W-Writing (n = 81)

Instructor interview results

The thematic analysis conducted on the instructor interview protocols identified a total of six themes on various aspects of student and instructor involvement in the assessment and learning activities based on the W-Writing module in the classroom: usefulness of W-Writing for students and instructors; students' and instructors' engagement in W-Writing activities; W-Writing's influence on classroom assessment practice; the value of W-Writing in academic writing instruction; the possibility of incorporating self- and peer assessment; and instructors' previous academic writing teaching experience. Due to space limitations, however, this section presents only the first two, directly related to the immediate content feedback presented during the revision stage. For reporting the results in subsequent sections, excerpts from the interviews, originally in Japanese, were translated into English.

Theme 1: Usefulness of W-Writing for students and instructors. The most prevalent theme identified in instructor utterances was W-Writing's usefulness for students and instructors. The instructors' comments were highly positive and discussed this issue mainly in terms of two subthemes. The first subtheme concerned the appropriateness of the instructional content (the video content) focusing on the five rules of summarization, on which the immediate feedback was based. Various positive comments were made by the instructors on the value of teaching these rules in general. Instructors A, C, and D commented on the critical importance of these rules in writing good summaries. As an illustrative example, Instructor A stated that the W-Writing package gave him an opportunity to "explicitly explain rules for summarization," such as Rule 2 (i.e., omit unnecessary details). He elaborated on the importance of teaching them in relation to the source text difficulty. While the other instructors (those teaching the intermediate sections) agreed that the text level was appropriate for their classes, Instructor A felt that the source texts might be quite easy for his advanced-level students. However, his comment below shows a clear distinction he made between these summarizing strategies and text comprehension abilities:

> …even if they might think the source texts are easy, they might not have opportunities to be explicitly taught the five rules…they are in the first year, so it is not good if they do not have a good grasp of such steps…it is a very important activity even if the source text is easy.

In addition, Instructor C pointed out the value of including Rule 5 (i.e., polish the summary) in particular, which emphasized the need for paraphrasing. Noting the students' inclination to seek an easy solution, Instructor C stated that this rule was "very good because it controls students' tendency to make it easy by just keeping cutting and pasting information [from the source text]."

A second subtheme concerned the appropriateness of the immediate feedback design. While numerous instructors' comments were related to this subtheme, three points are worth noting. The first concerned the W-Writing design feature that incorporates comparisons of the learner's performance against that of the Expert Summarizer (Feedback 1 and 2). Instructor A's comment on Feedback 1 highlights a substantive merit of this design as a teaching aid: "what information to select is a very important step… [but] it is sometimes difficult to teach how to do so explicitly." Meanwhile, Instructors A and B noted the clarity of the user interface (UI) design that allowed easy comparison for the learner: the Expert text underlines displayed along with the learner's text highlights in Feedback 1 (Instructor A), and the side-by-side presentation of the learner's outline and the expert's outline in Feedback 2 (Instructor B). A second point was the feedback design that demonstrates a model performance to the learner. Feedback 3 was mentioned as the case in point. More specifically, Instructor A found it to be "a very helpful scaffolding, as students have to internalize these rules to become able to pay attention to them even without a list.” Similarly, Instructor D noted that the list “articulates what to pay attention to in much detail … and students can actually experience what an expert summarizer would be thinking [while summarizing].”

Thirdly, Feedback 4, which highlighted exact copies of strings of three or more words, was considered helpful in encouraging students to play an active role in judging paraphrasing appropriateness. Instructors B and D endorsed the presentation of the three-word rule implemented in Feedback 4. For Instructor B, this came from her previous experience of receiving questions from students on a guideline as to how long a copy is permissible for other source-based writing assignments. Instructor D also commented that this rule “could function as a guideline suggesting the need to change [the text] at the sentence structure level.” Furthermore, Instructor A commended the instruction presented along with the highlighted strings: the student is the one to make the final decision as to whether a highlighted string needs to be paraphrased by examining each instance carefully.

In addition to the above-mentioned merits of the specific feedback types, the instructors commented on more general features of the W-Writing design as strengths. As advantages for students, Instructor B mentioned the ability of W-Writing to deliver detailed feedback to each student immediately, while Instructor C pointed out that the W-Writing feedback would shift students’ attention from form to content. Advantages for instructors, stated by Instructors B and D, concerned the availability of rich information about students’ performance. Instructor B reported that the amount of information available to instructors through W-Writing was “more than sufficient”. Meanwhile, Instructor D pointed out that the module that records not only multiple summary drafts but also data on learners’ responses to each feedback type allowed him to “picture how students proceeded in the learning process.”

While the instructors’ comments on the W-Writing feedback design were positive overall, two issues for further consideration were raised. One concerned the presentation of the three-word rule for paraphrasing presented in Feedback 4. Instructor C pointed out that the presentation of such a rule should not discourage the use of multi-word units (Read, 2023), which appear widely in academic discourse as they are considered to facilitate L2 writing development. The other came from Instructor D, who suggested the need to present the authors' names along with the source texts. He felt that the lack of information in the version of the W-Writing employed would make it difficult to draw students' attention to the need to acknowledge sources when summarizing. However, both instructors commented that these points could simply be explained to students as supplemental information when they use W-Writing in class.

Theme 2: Students' and instructors' engagement in W-Writing activities. During the interviews, all instructors stated that the two classes based on the W-Writing activities went smoothly, except for a few instances of technical errors requiring teaching assistants' support. Two subthemes were identified for this theme. The first concerned students' high engagement. Two instructors who observed students in person (Instructors B and C) reported that students were highly focused on the tasks during the classes. While the other two instructors whose classes met on Zoom had limited information in this regard, they also felt that their students engaged in the tasks well and completed them smoothly. Comments providing reasons for their observations varied and included content that interested students (Instructors A and C); ease of understanding the task at hand (Instructor B); change of pace, which gave students a break from other class content based on the textbook (Instructor D); and the novelty of the activity (Instructor C). This last point made by Instructor C appears to be related to his perception of the W-Writing feedback that "differs from the type of feedback that we [instructors] can offer." After reviewing the four types of W-Writing feedback with the interviewer during the interview, Instructor C noted:

> ...I think it is very good, and now I understand...the students were working [on the tasks] very hard until the end of class. No wonder it happened, because this much feedback was presented in that order, one by one. No wonder why they were so focused then.

While the instructors' comments above support their perceptions of students' task engagement, the second subtheme captured a potential drawback of the W-Writing design: lack of instructor engagement. In principle, W-Writing was designed for supplemental use both in and out of class. For this reason, the structure of the module is simple, enabling the student to complete all activities individually by following the directions on the screen. While Instructor A described it as a strength of the design promoting self-paced learning, Instructor B felt that this self-contained nature of the module made it difficult for her to engage in the students' learning process in class.

Because things went smoothly, with no questions from students, Instructor B reported feeling "as though [she] was just monitoring self-study sessions." As a remedy, she commented on adding a peer review activity to her class in the future, where students can exchange their summaries with each other and discuss the levels according to the analytic rubric (see Sawaki, 2020) that best describe their peers' summaries.

As a summary, Figure 2 presents a thematic map of the results of instructor interview. Theme 1 (the usefulness of W-Writing for students and instructors) was explained by the appropriateness of the instructional content and the feedback design perceived by the instructors. Theme 2 (students' and instructors' engagement in W-Writing activities) was intricately related to their perceptions of W-Writing's usefulness for students and instructors, identified as Theme 1. Theme 2 comprised two contrasting subthemes. One was students' seemingly high engagement in the activities, as perceived by the instructors, which was explained by various module design features such as content that interested students, the change of pace, and the novelty of the feedback. At the same time, however, these design features that contributed to the smooth administration of the W-Writing activities made it difficult for the instructors to become engaged in the students' learning process through the W-Writing activities.

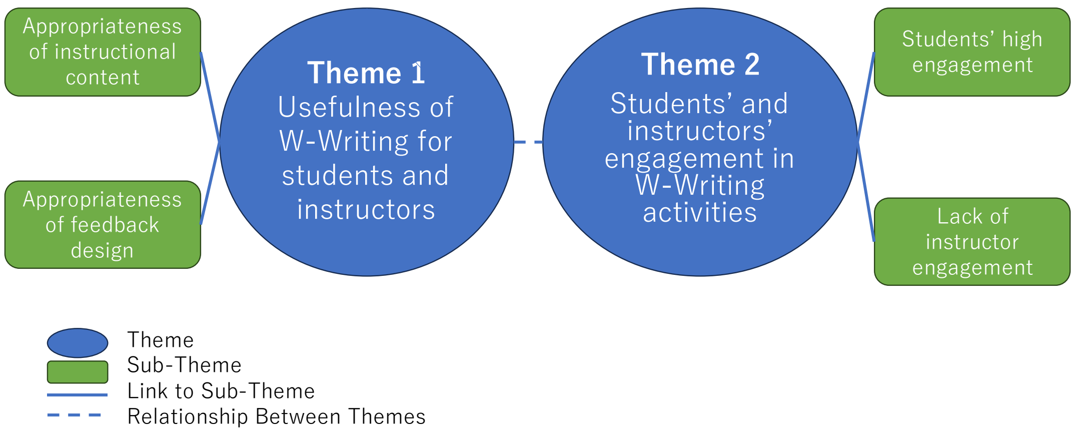

Figure 2. Instructor interviews: Thematic map

Discussion and conclusion

This mixed-methods study examined the efficacy of W-Writing, a formative assessment module on summary writing designed for supplemental use in undergraduate academic writing courses in Japan. The analysis of multiple data sources addressing two research questions suggested various benefits of utilizing technological capabilities brought by W-Writing for summary writing instruction in the new normal.

The results for Research Question 1 regarding Japanese undergraduate students' revision behaviors showed that students revised a majority (72%) of the sentences involved in their initial drafts. While surface changes to grammar and vocabulary were common, 69.9 percent of the total sentences revised included substantive changes corresponding to three W-Writing content feedback types. Moreover, the overall reduction in verbatim copying in the final draft suggested the students' attempts to revise text strings flagged as verbatim copies in the fourth feedback. At the same time, a detailed analysis of revision content showed the prevalence of relatively minor changes applied to specific details, while more major content revisions such as generating a topic sentence not included in the source text were relatively few.

The revision analysis results above require careful interpretation. The prevalence of relatively minor content changes to summary drafts as opposed to fewer major revisions is reasonable. Obviously, the number of specific details that can be manipulated in each source text is much larger than that of main ideas. Furthermore, previous summarization studies on the difficulty of macrorules (Brown & Day, 1983) suggest that those involving details (Rules 1 and 2 in this study) are easier than those concerning topic sentence use/generation and the synthesis of information across paragraphs (Rules 3 and 4). It is also reasonable to assume the need to adjust the amount of detail included to meet the summary length requirement. Thus, the relative frequency of different types of content revision observed in this study reflects an expected pattern. It should be kept in mind, however, that the current analysis captured differences between the initial and final drafts with the aim of illuminating incremental changes made to students' summaries upon content feedback provision. Thus, the absence of coding in a given category (e.g., representation of a topic sentence in the source text) does not necessarily mean that the target feature was also absent in a given student's summary drafts. It follows that a separate analysis examining the relationship between these revision operations and scores assigned to the initial and final drafts is required to explicate the degree to which these revisions resulted in score changes. Despite this limitation, it can be said that the W-Writing revision activities engaged students in meaningful summary writing operations.

Research Question 2 examined students' and instructors' perceptions of W-Writing feedback, which also offered supplemental information to interpret the findings on students' revision behaviors in relation to Research Question 1. Firstly, students' survey responses suggested a high degree of student endorsement of the usefulness of the four specific immediate feedback types and of W-Writing's usefulness for understanding the nature of summarization and learning summarization strategies and for future learning. These positive student attitudes toward the W-Writing feedback seem to explain their high degree of engagement in W-Writing feedback and meaningful revisions, improved quality being observed in Research Question 1. Secondly, similar to the students, the instructors showed positive views toward the W-Writing module in the thematic analysis. The most prominent theme identified in instructors' perceptions of W-Writing was its usefulness for students and instructors, explained by the appropriateness of instructional content and feedback design. This, in turn, was reflected in the instructors' interpretations of possible reasons for their students' seemingly high engagement in W-Writing activities.

The results for Research Question 2 offer valuable insights into stakeholders' perspectives on automated feedback in L2 writing, particularly those of instructors, which was identified as a current research gap (Koltovskaia, 2023). The instructors' positive views toward the W-Writing feedback revealed its various design strengths. One example that drew the instructors' attention was the feature enabling the learner to compare their own performance to that of the Expert Summarizer embedded in Feedback 1 and 2 (text underlines and outline, respectively). This result supports the feedback design principle to have the learner compare their work against an exemplar, which is considered to explicate expected performance standards (Carless & Boud, 2018; Orsmond et al., 2002). Another point worth noting is the generally high student engagement in W-Writing activities perceived by the instructors. This finding is in sharp contrast to previous L2 writing studies that showed low learner engagement in revision activities based on AWE feedback (e.g., time on task, number of drafts submitted; see Zhang & Hyland, 2018). The instructors' comments offer some clues concerning factors that may explain this finding, such as the students' interest in the content and the module design, where novel types of feedback on students' summary drafts were presented one by one during revisions. However, more detailed empirical data that offer insights into learners' perceptions are required to explore this issue more fully.

Despite the overall positive study results on stakeholders' views toward W-Writing, the reported difficulty of instructor engagement in students' learning activities merits further consideration. While the self-contained nature of the module with automated scaffolding for revisions may be suitable for independent learning, explicit instructor support may be required to help them use the formative assessment module as a springboard for conducting meaningful class activities. As noted by one of the instructors interviewed, a potential approach to addressing this issue is to combine the individual learning tasks based on W-Writing with additional formative assessment activities leading to collaborative classroom discourse, such as peer reviews of summary drafts. Providing instructors with such guidance for instructional planning might help them better integrate the formative assessment activities based on the W-Writing feedback into their writing classrooms.

All in all, a synthesis of the results for Research Questions 1 and 2 suggests that the present study offers initial empirical evidence of the overall efficacy of the four W-Writing feedback types from the perspectives of learner uptake of content feedback and stakeholders' perceptions of W-Writing usefulness. These findings also provide initial support for the W-Writing feedback design features inspired by the previous formative assessment and LOA literature. Such features include embedding the scaffolded revision activities in a larger learning task to promote learner uptake (Boud & Molloy, 2013); and actively involving the learner in the formative assessment process by incorporating a self-assessment checklist as a content feedback type (Andrade, 2010; Keppell & Carless, 2006).

However, it is premature to draw a hasty conclusion about W-Writing's efficacy based on this study alone due to its notable limitation: students' perceptions of the content feedback were reflected only in the survey results for W-Writing's usefulness. This, in turn, leads to another drawback: this study does not adequately address the multifaceted construct of learner feedback engagement. On the one hand, the findings of this study concerning the meaningful revision activities students engaged in and their seemingly high level of engagement in W-Writing activities reported by their instructors support their behavioral engagement in Zhang and Hyland's (2018) learner feedback engagement framework. On the other hand, other than students' high ratings for W-Writing's usefulness, this study offers no information relevant to the other two components in Zhang and Hyland's model: affective and cognitive engagement. Accordingly, it is possible, for instance, that the seemingly high engagement of the students in the revisions of their summaries observed above (a behavioral aspect of engagement) in fact reflects over-reliance on the content feedback, which indicates a low level of cognitive engagement (Koltovskaia, 2020). To thoroughly investigate such possibilities, future studies should explore learners' affect and strategy use underlying their observable revision behaviors and the high engagement in W-Writing activities observed in this study. Employing methods such as interviews and stimulated recall would paint a more comprehensive picture of the efficacy of this formative assessment module with immediate content feedback.

Last but not least, considering the flexibility of learning environments in the new normal, where various combinations of in-person learning and online learning are possible, an important issue for further exploration is whether the mode of instruction mediates AWE feedback efficacy. No such cross-mode comparisons were conducted between the Zoom and in-person classes in this study. This did not matter in this study because, across classes, all students completed all activities online individually, with virtually no interaction with the instructor or peers while on task. However, differences in the two modes may emerge as activities involving classroom interaction (e.g., peer assessment) increase. Future studies taking account of such mode differences would deepen our understanding of the intricate effects the learning environment has on how instructors and students perceive and utilize AWE feedback for teaching and learning.

Acknowledgments

The authors' special thanks go to all students and instructors who participated in this study and to our collaborators, Sayako Maswana and Hiroaki Yamada, for their help in data analysis for this study.

Author disclosures

This study was funded by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (B) (No. 20H01292, No. 24K00095).

The authors reported no potential conflict of interest.

The authors had the following roles respectively in conducting the research and writing the article: Yasuyo Sawaki – conceptualization, design of methodology, data collection, data analysis, writing original draft, revising, provision of research resources; Yutaka Ishii – design of methodology, data analysis, reviewing and editing of the article; Yoko Oi – design of methodology, data analysis, reviewing and editing of the article.

ORCID iDs

Yasuyo Sawaki https://orcid.org/0000-0001-6529-9929

Yutaka Ishii https://orcid.org/0000-0001-6054-8447

Yoko Oi https://orcid.org/0000-0003-4801-2273

References

Andrade, H. L. (2010). Student as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In H. L. Andrade, & G. J Cizek (Eds.), *Handbook of formative assessment* (pp. 90-105). Routledge.

Andrade, H. L., Brookhart, S. M., & Yu, E. C.Y. (2021). Classroom assessment as co-regulated learning; A systematic review. *Frontiers in Education*. 6. https://doi:10.3389feduc.2021.751168.

Bachman, L. F., & Damböck, B. (2018). *Language assessment for classroom teachers*. Oxford University Press.

Barrot, J. S. (2024). Trends in automated writing evaluation systems research for teaching, learning, and assessment: A bibliometric analysis. *Education and Information Technologies, 29*(6), 7155-7179.

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis*. (TOEFL iBT Research Report No. 14). ETS.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, policy & practice, 5*(1), 7-74.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698-712. https://doi.org/10.1080/02602938.2012.691462

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22*, 1-14. https://doi.org/10.1016/S0022-5371(83)80002-4

Carless, D. (2007). Learning-oriented assessment: Eonceptual bases and practical implications. *Innovations in Education and Teaching International, 44*(1), 57-66. https://doi.org/10.1080/14703290601081332

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Cerezo, L., Manchón, R.M., & Nicolás-Conesa, F. (2019). What do learners notice while processing written corrective feedback? A look at depth of processing via written languaging. In R.P. Leow (Ed.), *The Routledge handbook of second language research in classroom learning* (pp. 173–187). Routledge.

Choi, Y. & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic Bayesian networks. *Applied Sciences, 10*, 81-96. https://doi: 10.3390/app10228196

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Creswell, J. W. (2021). *A concise introduction to mixed methods research* (2nd ed.). Sage.

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies,* 1-53.

Fan, N., & Ma, Y. (2022). The effects of automated writing evaluation (AWE) feedback on students' English writing quality: A systematic literature review. *Language Teaching Research Quarterly, 28*, 53-73.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T.K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. E. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 68-88). Taylor & Francis.

Fu, Q.-K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: types, learning outcomes, and implications. *Computer Assisted Language Learning* https://doi.org/10.1080/09588221.2022.2033787

Hare, V.C., & Borchardt, K. M. (1984). Direct instruction of summarization skills. *Reading Research Quarterly, 20*, 62-78.

Hattie, J., & Timperlay, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Jang, E. E., & Sinclair, J. (2022). Diagnostic assessment in language classrooms. In G. Fulcher, & L. Harding (Eds.), *The Routledge handbook of language testing* (2nd ed., pp. 187-205), Routledge.

Keppell, M., & Carless, D. (2006). Learning-oriented assessment: A technology-based case study. *Assessment in Education: Principles, Policy & Practice, 13*(2), 179-191, https://doi.org/10.1080/09695940600 7039

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge Univ. Press.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394. https://doi.org/10.1037/0033-295X.85.5.363

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 100450.

Koltovskaia, S. (2023). Postsecondary L2 writing teachers' use and perceptions of Grammarly as a complement to their feedback. *ReCALL, 35*(3), 290-304. https://doi.org/10.1017/S0958344022000179

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Lam, D. M. K. (2021). Feedback as a learning-oriented assessment practice: Principles, opportunities, and challenges. In A. Gebril (Ed.), *Learning-oriented language assessment: Putting theory into practice* (pp. 85-106). Routledge.

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), 605-634. https://doi.org/10.1080/09588221.2020.1743323

Ngo, T.-T. N., Chen, H. H.-J., & Lai, K. K.-W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: a three-level meta-analysis. *Interactive learning environments, 32*(2), 727-744.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 27*(4), 309-323. https://doi.org/10.1080/0260293022000001337

Park, E.S., & Kim, O.Y. (2019). Learners' engagement with indirect written corrective feedback. Depth of processing and self-correction. In R.P. Leow (Ed.), *The Routledge handbook of second language research in classroom learning* (pp. 212–226). Routledge.

Read, J. (2023). Towards a new sophistication in vocabulary assessment. *Language Testing, 40*(1), 40-46. https://doi.org/10.1177/02655322221125698

Sawaki, Y. (2020). Developing summary content scoring criteria for university L2 writing instruction in Japan. In G. J. Ockey & B. Green (Eds.), *Another generation of fundamental considerations in language assessment: A festschrift in honor of Lyle F. Bachman* (pp. 153–171). New York: Springer.

Sawaki, Y., Ishii, Y., Yamada, H., & Tokunaga, T. (2024). Developing and validating an online module for formative assessment of summary writing with automated content feedback for EFL academic writing instruction. *Language Testing in Asia, 14,* Article 50. https://doi.org/10.1186/s40468-024-00325-w

Sadler, R. (1989) Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL, 36*(2), 187–209. https://doi.org/10.1017/S0958344023000265

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Turner, C., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari, & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255-273). De Gruyter.

Van Dijk, & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

Voss, E. (2021). The role of technology in learning-oriented assessment. In A. Gebril (Ed.), *Learning-oriented language assessment: Putting theory into practice* (pp. 207-224). Routledge.

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*(3), 333-362. https://doi.org/10.1207/s1532690xci2203_3

Wilson, J., Zhang, F., Palermo, C., Cordero, T. C., Myers, M. C., Eacker, H., Potter, A., & Coles, J. (2024). Predictors of middle school students' perceptions of automated writing evaluation. *Computers & Education, 211*, 104985.

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36*, 90-102.

Zhang, Z., & Hyland, K. (2023). The role of digital literacy in student engagement with automated writing evaluation (AWE) feedback on second language writing, *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2023.2256815