Automating individualized, process-focused writing instruction: A design-based research study

Keywords: keystroke logging, eye tracking, emerging writing technology, digital writing process intervention, design-based research

Abstract

Writing quality is dependent upon the organization and sequencing of cognitive processes during writing. College students need writing-strategy advice that is tailored to their individual needs and is cognizant of their already-established writing processes. However, there is an obstacle to providing such advice: Both writing instructors and the writers lack awareness of the moment-by-moment actions by which text was produced. This is because switching between the processes of defining the task, coming up with ideas, outputting text, evaluating, and revising is largely regulated implicitly.

To address this shortcoming, the present study uses a design-based research approach to develop and evaluate a minimally viable prototype of a system called “ProWrite” that uses novel biometric technology (concurrent keystroke logging and eye tracking) for providing real-time, individualized, automated, process-focused feedback to writers. This feedback is grounded in the analysis of each writer’s individual needs and is presented in the context of a learning cycle consisting of an initial diagnostic, an intervention assignment, and a final follow-up.

In two iterations, eight students used the system. Effects on student behavior were determined through direct analysis of biometric writing-process data before and after remediation and through changes in writing-process and written-product measures. Semi-structured interviews revealed that students generally considered the system useful, and they would try to use the newly learned strategies in their future writing experiences. The study demonstrated that individualized, real-time feedback informed by biometric technology can effectively modify writers’ processes when writing takes place.

1 Introduction

In postsecondary education, writing support is often provided in the form of required composition courses, disciplinary writing seminars, and (especially in the United States) in individual consultations in university writing centers. Individualized instruction is a staple of writing instruction at this level: Students routinely receive feedback about their texts. Nevertheless, many students struggle to become competent writers in their academic and professional careers (Duncheon and Tierney 2014).

A problem with the standard approach to postsecondary writing support is that students may not understand how to actually implement changes in their *writing process* that could improve their *written product*. Writing processes are the moment-by-moment actions that writers take over the course of producing a text (e.g., Breetvelt, van den Bergh, and Rijlaarsdam 1994; Vandermeulen, Leijten, and Van Waes, 2020). This includes component processes such as task analysis, planning, time management, brainstorming, translating ideas into language, typing, pausing, revising, and reviewing (Barkaoui 2019; Bowen and Van Waes 2020; Chukharev-Hudilainen et al. 2019; Galbraith and Baaijen 2019; Hayes and Flower 1980; Révész, Michel, and Lee 2019; Zhang et al. 2017). A written product, on the other hand, is the result of that writing process; i.e. the final text (e.g., Vandermeulen et al., 2020). There is clear evidence that the organization and sequencing of the component processes affects writing quality, with studies suggesting that a large proportion of the variance in writing quality can be attributed to the sequence in which processes are engaged during text composition (Breetvelt et al. 1994; Rijlaarsdam and Van den Bergh 2006).

Writers switch implicitly between component processes (Torrance 2015) but can, to some extent, consciously influence their coordination during writing (e.g., "I have to write a plan," or "I should write more before I read what I've written."). In primary and secondary education, the ability to take control of the writing process is often taught through "strategy-focused instruction" which offers substantial benefits over other forms of instruction (Graham and Harris 2003; Graham et al. 2012; Graham and Perin 2007). Strategy-focused instruction aims at making students aware of their own writing processes (i.e., what they do when). Meta-analyses of writing intervention research have found strategy-focused interventions—such as teaching strategies for pre-writing, idea planning, and goal-setting for productivity—to be successful for a wide variety of learners (Graham et al. 2012; Graham and Perin 2007; Rogers and Graham 2008), and more effective than, for example, grammar instruction and extra writing practice (Graham et al. 2012). By the end of secondary school, students have typically learned and automatized several of these writing-process strategies. However, when faced with novel and complex academic and professional writing tasks, university students are often unable to deploy their strategies effectively (Ranalli, Feng, and Chukharev-Hudilainen 2019). Therefore, any writing-process instruction that happens at the post-secondary level must take as a starting point each student's already-established strategies and attempt to improve those, instead of teaching writing strategies from scratch (Feng and Chukharev-Hudilainen 2017; Ranalli, Feng, and Chukharev-Hudilainen 2018, 2019).

While strategy-focused writing instruction (as well as earlier research into writing processes) relies on verbal protocols such as think-alouds and stimulated recalls, recent improvements in keystroke logging and eye-tracking technology (Chukharev-Hudilainen et al. 2019; Révész, Michel, and Lee 2019) permits a much more fine-grained analysis of the moment-by-moment actions taken during text production, including those actions that are not explicitly controlled by the writer and may not reach their awareness. Indices based on such analyses have been found to predict text quality. Baaijen and Galbraith (2018), for example, investigated a set of 11 keystroke-based variables, such as percentages of linear transitions from one unit (word or sentence) to the next and mean pause durations between sentences. The researchers used principal component analysis to aggregate these indices into two composite measures, one capturing revisions of global text structure and the other showing the extent to which sentence production is pre-planned or spontaneous. In the sample of 84 university students, writers who produced sentences more spontaneously also tended to produce higher-quality texts, while the relationship between global revisions and text quality varied based on the type of global planning that the writers did before starting to output their text. The finding that keystroke measures predict writing quality suggests that providing students with feedback that is based on such measures might help them improve the quality of their texts.

The potential value of individualized instruction based on keystroke measures has been investigated by Bowen, Thomas, and Vandermeulen (2022) who used keystroke logging software (specifically Inputlog; Leijten and Van Waes 2013) to provide process-focused feedback to university students. Three days after completing a draft of their writing assignments, participants were presented with various statistics derived from the participants' keystroke logs (such as the percentage of time during the writing session when they were not actively typing, or the percentage of text they produced during a particular portion of their writing time-course) shown alongside benchmark statistics obtained from a set of high-quality essays. This type of feedback was found to increase the use and awareness of metacognitive writing strategies, particularly ideational planning. Participants also described this feedback as engaging and interesting. However, Vandermeulen (2020) reported that students found the interpretation of this feedback difficult; students also stated that they were not confident in their ability to improve their writing in response to the feedback and could not recall most of the feedback in subsequent writing sessions. Addressing this concern, Vandermeulen et al. (2020) developed a new function in Inputlog that was designed to facilitate the provision of process-focused feedback by writing tutors to students. Their study highlighted the important role of the writing tutor in selecting process variables of interest to individual students and personalizing process feedback.

A different approach to providing process-based feedback was proposed by Ranalli, Feng, and Chukharev-Hudilainen (2018). Instead of helping students adjust their writing processes so that they would lean toward process-measure benchmarks obtained from high-quality texts, researchers manually and qualitatively examined replays and visualizations of writing processes and used their intuition as writing instructors to identify behaviors within the writing process that might plausibly lead to specific deficits within the final text in the specific writer. For example, a writer who produces short texts that go off-topic might be struggling because they frequently pause mid-sentence to look up words in a dictionary. Importantly, and in contrast to the above approach, frequent mid-sentence pauses are not considered an a-priori intervention target: it may well be the case that a different writer might be highly effective while also pausing mid-sentence. However, mid-sentence pauses in a writer with particular product deficits may trigger feedback if the analyst believes, based on their domain knowledge, that this pausing behavior might be causally related to the shortcomings of the text. In this case, it would be reasonable to suggest to the writer that they defer dictionary look-ups until the end of their writing process. In the reported case-study series, two students completed a series of four argumentative writing assignments. Upon the completion of a writing session, the participants met with one of the researchers to discuss their writing process while viewing animated, keystroke-by-keystroke playbacks of their writing process overlaid with a semi-transparent gaze marker showing how the writer's attention shifted back and forth between the current word and previous words in the text they had written. As a result, the researchers suggested individually tailored writing strategies to the students. The students reacted favorably to the suggested writing strategies. However, providing this type of instruction required time-consuming manual labor and thus would be impossible to scale. Additionally, learners forgot the advice they were given and failed to implement the suggested writing strategies in the subsequent writing sessions, a shortcoming that could be resolved through real-time feedback provision (as opposed to providing feedback after the session was complete).

This latter shortcoming, in fact, is characteristic of all instructional approaches reviewed so far. Providing feedback to learners after their writing session, rather than concurrently during the writing session, creates a time gap between the suboptimal behavior and the feedback that aims at its modification. As an early step toward closing this gap, following a proof-of-concept pilot by Feng and Chukharev-Hudilainen (2017), Dux Speltz and Chukharev-Hudilainen (2021) investigated whether real-time process feedback provision might be beneficial. In their study that followed a within-participant design, 20 native-English-speaking undergraduate students wrote two short essays each under a strict time limit of 10 minutes. One of the essays was randomly assigned to an experimental condition wherein the opacity of the text on the screen was reduced incrementally whenever the writer paused. This "disappearing text" was provided as a form of feedback to encourage the participant to avoid any interruptions and carry on the composition process. In the controlled condition, the participants were informed about the benefits of fluent text production given the time constraint, but were not provided any real-time feedback during their writing session. In the experimental condition, the quality of participants' texts, as assessed through a holistic rubric, improved, while their accuracy suffered and complexity remained unchanged. The participants reported feeling focused, intentional, and motivated in their writing. Importantly, the study demonstrated that writers responded to the real-time feedback as intended (i.e. by increasing the fluency of their writing). When controlling for the length of the texts, there was no significant difference in the text quality between the control and the experimental conditions, indicating that the intervention was most helpful for increasing text length which in turn led to higher text quality. This finding has also been supported by previous research which has shown text length to be a stable indicator of overall text quality (e.g., Bennett et al. 2020). To our knowledge, this is the only study that has provided real-time process feedback.

Drawing on the findings and the limitations of previous research, this paper aims to explore the feasibility of designing a computational system that will provide actionable process-focused feedback to university-level students of writing based on automatic analysis of the student's keystroke and eye movement data. The characteristics of the proposed system will be outlined in the following section.

2 The Present Approach

In the present study, we used a design-based research approach to iteratively develop "ProWrite," a prototype system for delivering writing-process feedback. ProWrite is based on CyWrite (Chukharev-Hudilainen et al. 2019; Chukharev-Hudilainen and Saricaoglu 2016), an open-source web-based text editor that includes built-in functionality to capture synchronized keystroke and eye-movement logs which can later be viewed as a replay of the writing process, visualized as interactive graphs, or exported for analysis with external programs. In the ProWrite system, the following additional functionality has been implemented: Keystroke latencies are analyzed in real time to identify locations where the writer pauses (for example, word-initial vs. within-word); eye fixations are analyzed during each pause to identify the locations in the text the writer is paying attention to (for example, when they reread text); and revisions are analyzed to determine whether the writer is producing text or deleting a text segment. As a result of these analyses, the ProWrite system can be configured to display real-time scaffolding prompts to guide the user in their writing process. The ProWrite system affords a three-stage learning cycle comprising the following:

1. **A diagnostic session** where the writer composes a text without any advice about how they should modify their writing processes while recording data from keystrokes and eye movements (the goal here is to capture the writer's baseline writing processes)
2. **An intervention session** where the writer is presented with an analysis of their writing processes, guided by a human consultant to determine an appropriate remediation plan, and then composes a new text while enacting this plan with real-time assistance from an automatic scaffolding system
3. **A follow-up session** under conditions as in (1) without advice or scaffolding with the goal to assess the effect on the writer's behavior and text quality beyond the intervention session in (2).

In two design iterations, we developed and evaluated a prototype of the system and gathered insights for improving the learning cycle. Where there is evidence that participants understand and embrace a specific remediation plan and are able to adapt their behavior accordingly, the remediation plan will be passed forward to the next iteration. Ultimately, this design-based process is expected to yield a system that implements a range of diagnostic rules that are plausibly beneficial, and that can be understood and implemented by learners, even without scaffolding. However, evaluating the benefits of the system is beyond the scope of the present paper. Here, the focus is squarely on establishing the feasibility of a system with the above-mentioned characteristics.

The remediation plans selected for the first two iterations reported in this article were designed to encourage writing-process behaviors meeting two criteria: (1) the target behaviors have been identified as beneficial for improving text quality in previous research, and (2) in an existing inhouse dataset of writing processes (similar tasks and participant characteristics, 40 writing-process recordings from 20 participants), we observed plausible causal links between writing-process behaviors (identified as candidates for remediation) and written-product deficits. These observations were made informally by the first and the last authors drawing on their pedagogical intuitions. The remediation plans were as follows:

1. "Do not edit": postpone revisions beyond the leading edge until a full draft of the text has been produced. This remediation plan was considered for a group of participants who were seemingly "stuck" and not producing a lot of text because they were engaging in repeated revisions of the same text span, usually at a sentence level, with these repeated revisions not leading to any observable improvement.
2. "Pause sentence-initially": take a moment to plan out the sentence at hand before beginning to type it. Qualitative observations identified participants who rarely paused sentence-initially and instead paused mostly mid-word or mid-clause, possibly to think about how they wanted to finish the sentence. At times, this behavior seemed to lead to incohesive sentences and a failure to connect ideas.
3. "Revise periodically": upon completing each paragraph, take a moment to re-read what you have written and make revisions. This plan was considered for participants who exhibited a highly linear writing behavior, with few if any recursions of the inscription point from the leading edge of the text. Previous studies have found that revision behavior is related to text quality (Barkaoui 2016; Vandermeulen et al. 2020), and Baaijen and Galbraith (2018) found that higher levels of global revision, specifically, were associated with better text quality. This remediation plan aims to encourage such global revision by asking writers to read their already-written text periodically to ensure that each paragraph fits the essay's overall goals and structure.
4. "Write linearly": write at the leading edge until a complete draft has been written. In contrast to the previous remediation plan, this one was considered for those participants whose writing behavior was highly non-linear: they jumped around in their text during the writing process, sometimes leaving the middle of the word or sentence at the leading edge to add entire sentences to previous paragraphs.

In general, these remediation plans capture aspects of (1) revision behavior and (2) pausing and fluency, which have continuously been identified as important aspects of the writing process (e.g., Vandermeulen et al. 2020; Bowen et al. 2022). The optimal strategy may, in fact, be a combination of two or more remediation plans. The focus of the present study, however, was not on determining whether such optimal strategy exists, or whether the proposed remediation plans are effective at improving writing quality. Instead, our focus was on exploring whether these writing features can be manipulated independently by targeting one at a time during the writing process.

3 Design Iterations

This article describes the first steps in developing an automated system that provides writing instruction by directly modifying the writing process through the development of individualized process remediation plans and their real-time scaffolding. Because this type of system has not been proposed previously, a design-based research (DBR) approach was adopted with multiple design iterations each involving a limited number of system users. In the DBR approach, each iteration's findings inform changes that are made to the system in the next iteration (Brown 1992; Collins 1992). Importantly, the goal of DBR is to "develop and refine interventions based on the results of studies" (Feng 2015, 47) rather than setting out to verify a predetermined hypothesis as is the case for predictive research. DBR studies are situated in a real educational context, where they focus on the iterative design and testing of an intervention (typically using a mixed-methods approach) in collaborative partnership between researchers and practitioners. DBR approaches are being increasingly used in educational settings, and they are especially beneficial for studies aiming to develop effective and robust technological interventions that have a practical impact on practice (Anderson and Shattuck 2012; Reeves and McKenney 2013). The next section presents the methods and findings for each design iteration in the present study.

3.1 Participants

The participants in this study were 13 native-English-speaking undergraduate students (mean age: 20 years; range: 18–22; 12 female, 1 male) at a large Midwestern research university in the United States. Eight participants participated in each design iteration. Three participants from Iteration 1 also participated in Iteration 2 to allow for within-participant comparisons. All participants were compensated with electronic gift cards for their participation. This study was reviewed and considered exempt by the university's institutional review board. The participants provided their written informed consent to participate in the study.

3.2 Iteration 1

Iteration 1 prioritized developing procedures for the analysis of keystroke logging and eye-tracking data, establishing a comprehensible and usable structure for specifying remediation plans, and developing mechanisms for deriving scaffolding feedback from the remediation plans. This iteration took place July–September 2021 and was guided by the following research questions:

RQ1. To what extent do learners adapt their writing processes in line with the selected remediation plan and the real-time scaffolded feedback?

RQ2. What is the experience of learners using the ProWrite system?

3.2.1 Methods

Eight participants experienced the system and learning cycle during this iteration. In this iteration, the eye-tracking hardware was an industry-standard, research-grade eye-tracking system—an SR Research EyeLink 1000 Plus Remote (set to 500 Hz sampling rate, providing less than a 1° spatial error).

3.2.1.1 The Diagnostic Session

In the Spring 2021 semester (7 participants) or Fall 2021 semester (1 participant), participants first attended a 90-minute diagnostic session in which they composed two essays for up to 35 minutes each in response to the following prompts (counterbalanced for order):

A. "Some people have said that finding and implementing green technologies, such as wind or solar power, should be the focus of our efforts to avert climate crisis. To what extent do you agree or disagree with this statement?"

B. "Some people have argued that animals should be given similar rights to humans. To what extent do you agree or disagree with this statement?"

These prompts were selected because they encouraged writers to use higher-level writing skills (such as developing claims, incorporating evidence, and organizing ideas) while still being accessible for college students in the United States without requiring specialized knowledge or use of sources. During this diagnostic session, writers were asked to write a five-paragraph academic essay to the best of their ability. Upon completing the first essay, participants were instructed to take a break for up to 10 minutes before beginning the next essay. We then used these diagnostic essays to determine the remediation plan that would be most relevant to each writer based on their process behavior.

3.2.1.2 Selection of the Remediation Plan

Two remediation plans were selected for this iteration based on the most common process behaviors that appeared to lead to issues in the final texts of the 40 manually analyzed sessions: “Do not edit” and “Pause sentence-initially.” During this iteration, similarly to Ranalli and colleagues (2018), an expert decision about the optimal remediation plan was made in each case by the researchers. In the decision-making process, the authors and a trained research assistant manually reviewed visual replays of the writing processes, and noted features such as the presence/absence of final review, episodes of inscription, time that writing begins, time between the first read-through of the prompt and the beginning of writing, frequency/volume of deletions, frequency of evaluation and/or reading of the prompt, overlapping of processes, frequency/duration of pauses, location of longest pauses, and presence/absence of sentence-initial pauses. Decisions for remediation plans were made based on these visual analyses because in this iteration, no rules had yet been established to determine thresholds for triggering each remediation plan. By analyzing the visual replays, it was possible for researchers to determine when certain process behaviors appeared to be problematic based on their expert experience. Table 1 summarizes the patterns of process behavior that led to the selection of each remediation plan.

The manual process analysis can be illustrated by examining the diagnostic writing session from one participant. Participant 1 did not have a final review session during her first diagnostic essay, and she had a limited final review session that seemed incomplete during her second diagnostic essay. Both diagnostic essays had an episode of inscription, one of which was around two minutes long. In both essays, she began writing after 20–40 into her sessions, with one second after her first read-through of the prompt in one session and 31 seconds after her first read-through of the prompt in the other session. She had several episodes of deletion, some of which were large deletions. She paused before some paragraphs, at times to re-read the prompt, and had some sentence-initial pauses. With this pattern of behavior, it was determined that she would be a good candidate for the “Do not edit” remediation plan (see Table 1). Using this manual process analysis, four writers were assigned the strategy of “Do not edit” and four were assigned “Pause sentence-initially.”

3.2.1.3 Text Quality

Quality of participants’ texts was assessed manually. We first consulted with assessment experts for advice about creating a writing rubric for the purpose of assessing texts in the present study. Upon conferring with several writing pedagogy experts and practitioners, it was determined that text quality for all essays would be assessed using an analytic rubric with the following measures: (1) the quality of the introduction, (2) the quality of the conclusion, (3) the essay’s adherence to the prompt, (4) the ability of each body paragraph to stay focused on a single main idea, and (5) the quality of the transitions between paragraphs. A four-point analytic scale for each of these categories was used. The paragraph-level measures were developed in order to allow for paragraph-level analyses of both the product and the process. The rubric can be found in Appendix A.

Because identical prompts were used for all participants in the intervention and follow-up sessions, these ratings were not used to compare the written quality of texts in different session types (as any difference between texts in different session types could be attributed to the prompt). Instead, these scores were only used to provide the researchers with areas of concern in a participant’s written product. Two trained undergraduate research assistants rated each text collaboratively; it was determined that independent ratings were not required for calculating inter-rater reliability as the rubric was only being piloted for a limited written-product-quality overview during this iteration.

3.2.1.4 The Intervention Session

For each remediation plan, we developed an automatic feedback provision mechanism as follows: For the "Do not edit" plan, a feedback message would appear after the participant started deleting characters in their text beyond the word they are currently typing. For the "Pause sentence-initially" plan, a feedback message would appear at the start of each sentence.

Participants scheduled a 1-hour session to return for a one-on-one intervention session with a researcher (one of the authors or a trained undergraduate research assistant) during the Fall 2021 semester. The average length of time since the diagnostic session was 104 days (min = 21 days, max = 157 days). To begin this session, the researcher presented the participant with a writing-process recording of one of the participant's diagnostic essays and first asked the participant to briefly become reacquainted with what they had written. Next, the researcher asked the participant how they felt about their essay and whether they felt the essay had any issues. After listening to what the participant identified as potential areas of concern, the researcher pointed out a few product issues identified by the manual product analysis. The researcher then explained what had been documented about the session during the manual process analysis and presented the participant with brief replays—which included a keystroke-by-keystroke playback of the writing process and a semi-transparent gaze marker showing how their eyes moved during this time—of those parts of their process, explaining what the manual review had identified as potentially problematic or otherwise unique process behaviors. These replays illustrated the process concerns and tied the writer's cognitive representation of the problem to their memory of the problematic actions they took during their diagnostic session. Explanations of process behavior were presented neutrally and factually, with the researcher purposefully avoiding any claims that the participant's writing process was inherently wrong. Next, the researcher introduced the proposed remediation plan and explained in detail precisely how the intervention would take place on the ProWrite system. Importantly, this remediation plan was introduced as a way to experience a new writing strategy. The intervention was suggested by the researcher as something that could potentially improve the participant's essay quality if the participant agreed that it could be worthwhile. Through this approach, the participant took agency of the remediation plan and willingly agreed to try the intervention for the writing session.

Upon agreeing to try the proposed remediation plan, participants composed a new text, this time only completing one essay for up to 35 minutes. The writing prompt was as follows: "Science should aim to discover the truth about the world, without concern for practical application or wealth creation. To what extent do you agree or disagree with this statement? Try to support your arguments with, for example, your knowledge of scientific evidence, specific examples from your own experience, or your observations and reading." During this session, real-time scaffolding feedback appeared on the screen to remind the participant to adhere to the remediation plan. For the "Do not edit" intervention, a pop-up box with the words "Do not edit your text. Just keep writing." appeared on the screen whenever a participant began deleting beyond the word they had just typed (see Figure 1a). For the "Pause sentence-initially" intervention, a pop-up box simply reading "Think!" appeared upon the completion of each sentence (see Figure 1b). These reminders were formulated with the goal of concisely reminding participants of the core goals of each remediation plan. The yellow circle in Figure 1b indicates where the writer was looking at the screen.

Immediately after the participant completed their essay, they participated in a brief informal interview session to gauge their opinions about the remediation plan and the scaffolded real-time feedback in the form of pop-up boxes. As a starting point, the researcher asked the participants to share how they felt about the intervention, whether they felt the feedback was distracting, and what their thought process was while writing with the remediation plan in mind. Follow-up questions were asked to clarify responses as needed.

3.2.1.5 The Follow-up Session

Finally, participants returned once more to complete another non-scaffolded writing task in order to determine whether the participants retained the learning goals of the remediation plan. These sessions also took place during the Fall 2021 semester. The average time since the intervention session was 12 days (min = 4 days, max = 22 days). Participants wrote one more five-paragraph argumentative essay for up to 35 minutes and were instructed that they could write using any strategies they would like. The writing prompt was as follows:

> "As we acquire more knowledge, things do not become more comprehensible, but more complex and mysterious. To what extent do you agree or disagree with this statement? Try to support your arguments with, for example, your knowledge of scientific evidence, specific examples from your own experience, or your observations and reading."

Upon completion of this essay, participants completed one more informal interview with the researcher. To gauge how well the participant remembered and understood the remediation plan from last time, participants were asked, "Do you remember the game plan from your last session?" Then, to establish whether they consciously chose to utilize the same remediation plan as last time, they were asked, "Do you think you tried to use this strategy during your session today?"

3.2.2 Findings

To address the first research question, effects on student behavior were determined through manual analysis of biometric writing-process data before and after remediation. To address the second research question, participants' responses to the informal post-intervention interview were analyzed descriptively.

3.2.2.1 RQ1 findings: Do not edit

Manual analysis of writing-process data for the participants who received the "Do not edit" remediation plan consisted of visually analyzing graphs that display editing behavior. Figures 2a–2d show the process graphs representing editing behavior for Participants 1, 2, 5, and 6, the recipients of the "Do not edit" remediation plan, during their diagnostic, intervention, and follow-up sessions. The horizontal axis represents elapsed time during the writing process (in minutes), and the vertical axis represents the changing engagement in various writing processes, with metrics averaged within a sliding five-second window. The gray line represents text production, and the red line represents text deletion, with both lines scaled based on the participant's typing speed. Vertical spikes in the red lines, therefore, represent periods of increased deleting activity.

A qualitative inspection of Figures 2a–2d revealed that participants made substantial changes to their editing behavior as observed in the diagnostic session during their intervention sessions, which demonstrates that they were able to adhere to the remediation plan effectively. Whereas the two diagnostic sessions for all four of these participants show large episodes of deletions scattered throughout the writing process and minor deletions almost constantly during the writing process, the intervention sessions show less frequent editing behavior. Participant 6's periods of editing activity during the intervention came only during a period of final evaluation, which was allowed as part of the remediation plan. Participants varied in terms of whether they adhered to the remediation plan in the follow-up sessions. For example, Participant 1's follow-up session process was quite similar to the intervention session, whereas Participant 2's follow-up session process reverted back to a process similar to that of her diagnostic sessions.

3.2.2.2 RQ1 findings: Pause sentence-initially

Writing processes for participants who received the "Pause sentence-initially" remediation plan were also analyzed manually by the researchers using visualizations of pause durations. These visualizations are presented in the form of red highlights overlaid on the text that was produced. Dark red highlighting indicates a longer pause that occurred during the writing process. Crossed-out text indicates text was subsequently deleted. Figure 3 presents the visual interface for one paragraph of Participant 4's diagnostic, intervention, and follow-up sessions.

As shown in Figure 3, Participant 4's diagnostic sessions included inconsistent and problematic pausing behavior. She frequently paused within sentences, demonstrating that she may not have had a clear idea of where her sentence was going before she started to type it. In the intervention session, she successfully adhered to the remediation plan by pausing at the beginning of sentences. In fact, in both the intervention session and the follow-up session, all of the longest pauses (indicated by the darkest red shading) appeared sentence-initially. An interesting finding for this participant is that her diagnostic sessions included a substantial amount of deleting behavior, but there were very few deletions in her intervention and follow-up sessions. This demonstrates the potential of the "Pause sentence-initially" remediation plan to reduce text revision, which is consistent with Baaijen and colleagues' findings that controlled sentence production with long sentence-initial pauses leads to reduced deleting behavior (Baaijen et al. 2012; Baaijen and Galbraith 2018). Participant 3 followed a similar pattern of behavior as Participant 4: her diagnostic sessions showed excessive pausing behavior within-sentences, her intervention session showed successful adherence to the remediation plan, and her follow-up session showed that she continued to implement the remediation plan even without scaffolding.

Participant 7, on the other hand, only partially continued to use the "Pause sentence-initially" strategy during the follow-up session, even though she successfully adhered to the remediation plan during the intervention session. Participant 7 continued to occasionally pause mid-sentence, unlike Participants 3 and 4 who no longer paused mid-sentence when intentionally pausing before each sentence. This could suggest that the participant did not dedicate enough time to prepare what she was going to say before she started writing the sentence (see Figure 4 for screenshots of one paragraph out of each of Participant 7's texts).

Finally, Participant 8 was unique from the other three participants in that he was assigned the "Pause sentence-initially" remediation plan due to how little he paused during the diagnostic sessions. His intervention and follow-up sessions show more pauses at sentence-initial positions (see Figure 5 for screenshots of one paragraph out of each of Participant 8's texts).

2.2.2.3 RQ2 findings

With regard to the second research question, participants generally considered the system to be useful for three main reasons: namely, they experienced their own writing process as more intentional, it successfully reminded them of the assigned remediation plan, and it increased their awareness of the writing process. All eight participants used positive adjectives such as "helpful," "nice," and "good" to describe the intervention. For example, Participant 1 received the "Do not edit" intervention, and she said that she felt good about the intervention because it "made me think about what I was typing […] and I think it made me think of more useful words." In this way, because she was instructed not to delete, she planned her words carefully as she was typing and selected words that she deemed to be most useful after careful consideration.

The pop-up boxes were also generally well-received by the participants, although three participants noted that they could be distracting. One participant who received the "Do not edit" intervention said, "[The pop-ups were] a good reminder just because I definitely delete more than I probably should" (Participant 5). However, this participant also noted that she felt that she was not adhering to the goals of the intervention perfectly. She noted, "I definitely still did delete some, so it might be good to practice more," and she continued to say that the process was quite unnatural for her because she wanted to delete when she realized she did not like something she had written.

The "Think!" message for the "Pause sentence-initially" intervention triggered the following thought process for Participant 7: "It was definitely nice to have that pop-up […] [When I saw the pop-up,] I took any ideas that I had coming in and tried to form them right there instead of forming them as I went." Additionally, the intervention caused Participant 7 to realize when she paused in the middle of a sentence: "[The intervention] definitely helped me notice when exactly in the sentence I stopped because I was thinking of a word." In other words, even when she was in the middle of a sentence, she thought about the pop-up box that she had received at the start of the sentence, and she realized that she was pausing in the middle of the sentence to mentally search for a word. This self-awareness could allow her to self-correct pausing behavior even when no pop-up messages were provided. Therefore, although the pop-up message may have been distracting, it achieved its primary purpose of reminding the participant of their remediation plan.

Participant 2 had a similar experience. She was more aware of her writing process without necessarily relying on the pop-up boxes. She was assigned the "Do not edit" intervention. Thus, she could avoid the pop-up box by not deleting past the word she was currently typing. She noted that she did not receive many pop-up reminders because she had mentally gamified the system: "When I saw [a pop-up box], I thought, 'Darn, I thought I could get away with it!'" She would only start typing the next sentence after carefully considering and planning it out to avoid subsequent editing of the text.

All eight participants could clearly and accurately describe the intervention they had been assigned when returning for their follow-up session. Seven of them noted that they thought about trying to implement the intervention strategy during their follow-up session even though there was no explicit instruction to do so. For instance, Participant 5 confirmed that she tried to adhere to the intervention that she was assigned to during the follow-up session. She noted this was harder than expected. When asked why she tried to follow the remediation plan even though there was no instruction to do so, she reflected, "I remembered it from last time […] I thought it was useful just getting my information on the page instead of second-guessing myself." Similarly, when Participant 4 was asked whether she attempted to follow the "Pause sentence-initially" strategy, she explained, "I did because last session I was a little stressed because it kept popping up with the message, but this time I just had it in my mind. So I just thought 'Pause, think about what I want to say next.'" Participant 8 was the only participant who claimed that he did not try to adhere to the intervention strategy during the follow-up session, even though he said that he "felt good about it" during the intervention session.

2.2.3 Reflection

The primary goals of the first iteration were (1) to establish a system for providing automated, real-time, process-focused feedback to learners, (2) to determine how learners changed their writing process in line with the remediation plan when receiving this feedback (RQ1), and (3) to gain insight into learners' experience using this system (RQ2). Two remediation plans were developed and implemented in the ProWrite system with automated scaffolding: "Do not edit" and "Pause sentence-initially."

The first research question—how learners changed their writing process in line with the remediation plan—was addressed by comparing writing process behavior during the initial diagnostic session and during the intervention session. The participants' writing process changed during the intervention session in correspondence to the assigned remediation plan. Participants who received the "Do not edit" remediation plan avoided editing during the intervention sessions, and participants who received the "Pause sentence-initially" remediation plan paused substantially more sentence-initially and less at mid-sentence locations in the intervention sessions. This demonstrates that participants were able to successfully change their writing-process behaviors in accordance with the assigned remediation plan supported by real-time scaffolding. However, participants varied in terms of whether they continued to enact the remediation plan during follow-up sessions. This could be due to one of two reasons: (1) the participants tried and failed to enact the remediation plans without automated scaffolding, or (2) the participants did not try to enact the remediation plans when they were not explicitly told to do so. Therefore, in Iteration 2, the researchers told participants that they believed it would be beneficial to continue to enact the remediation plan, and participants were asked about whether they intended to do so in the post-follow-up interviews.

The second research question—which concerned learners' experience with the system—was addressed by conducting semi-structured interviews following both intervention and follow-up sessions. Participants overwhelmingly supported the usefulness of the intervention, and they appreciated the pop-up boxes as reminders of the remediation plans, even though a few participants considered them distracting. This is not unexpected because the pop-ups were deliberately designed to be distracting to be noticed by the learner and remind them of the remediation plan. To alleviate concerns that distraction might be preventing learners from creating the best text possible, the researchers warned the participants in the second iteration that the pop-ups may be distracting and that this was intentional.

The research team reflected on the methods and findings from Iteration 1 and formulated four goals for Iteration 2 as part of the DBR improvement process:

1. **Moving to a deployable eye-tracker**. In Iteration 1, an SR Research EyeLink 1000 was used to test the ProWrite system with the best research-grade eye-tracking equipment available. For Iteration 2, the research team decided to use a GazePoint GP3 HD eye-tracker as a step toward developing a system that can be used in non-laboratory settings.
2. **Increasing the number of remediation plans.** Iteration 1 demonstrated that participants were able to adhere to remediation plans. Two additional remediation plans were added to the system for Iteration 2.
3. **Semi-automating the selection of remediation plans.** In Iteration 1, remediation plans were selected using the manual review of writing-process visualizations. For Iteration 2, the writing-process components were summarized as numerical metrics that could be automatically extracted from the biometric data. We describe these metrics in the following section.
4. **Adjusting the phrasing of the remediation plans and their associated automated scaffolding.** To provide positive and actionable advice, which is standard practice for developing interventions in young children and also relevant for the present study[1], the research team decided to adjust the phrasing of the remediation plan "Do not edit" to "Commit to finishing your sentence."

[1] We appreciate Dr. Carolyn Richie's advice that led to this design modification.

3.3 Iteration 2

The following research questions were derived to guide Iteration 2:

RQ1: How can automatic process metrics be used for the selection of the remediation plan?
RQ2: How do the writing process and written product change based on the remediation plan?
RQ3: How do users experience the new version of the ProWrite system?
RQ4: To what extent do returning participants sustain previously taught process modifications?

3.3.1 Methods

As in Iteration 1, eight participants experienced the system and learning cycle during this iteration. This iteration took place in February and March 2022. An improved version of the ProWrite system was used in this iteration, but the implementation of keystroke and eye movement capture was unchanged. A GazePoint GP3 HD eye-tracker (150 Hz) was used in Iteration 2. Although the GazePoint GP3 HD system had a poorer temporal and spatial resolution than the SR Research EyeLink 1000 Plus, pilot work prior to Iteration 2 suggested that accuracy was sufficient to capture eye movement behavior (e.g. reading during writing pauses).

3.3.1.1 Summary of Sessions and Prompts

As in Iteration 1, participants attended three sessions led by a researcher (the authors or a trained undergraduate research assistant). In Iteration 2, participants were asked to write one essay per session (participants in iteration 1 wrote two diagnostic essays). Three prompts similar to those used in Iteration 1 were selected with the intention to elicit the same higher-level writing skills while discussing topics accessible for college students in the United States. The writing instructions for the diagnostic session were the same as in Iteration 1. Prompts were counterbalanced and included the following:

A. “Social media and streaming algorithms are responsible for recommending diverse items to avoid creating an echo chamber. To what extent do you agree or disagree with this statement?”
B. “Success in education is influenced more by the student's home life and training as a child than by the quality and effectiveness of the education program. To what extent do you agree or disagree?”
C. “Society should make efforts to save endangered species only if the potential extinction of those species is the result of human activities. To what extent do you agree or disagree with this statement?"

3.3.1.2 Text Quality

Three raters assessed the quality of the texts using the rubric from Appendix A. Each essay was assessed independently by two raters. Exact simple percentage agreement was 44%, but adjacent agreement (within 1 point) was 86%, which was deemed acceptable for the purpose of the present study, but was noted as a limitation that needs to be addressed for future larger-scale implementations.

3.3.1.3 Selection of the Remediation Plan

In line with the second improvement goal from Iteration 1, two new remediation plans were introduced: “Write linearly” and “Revise Periodically.” The operationalisation of linearity measures requires clarification: Linearity of writing was operationalized by identifying “blocks” of text that participants produced continuously. Continuous production was operationalized as follows: A “block” was defined as a stretch of text that was produced either (1) without cursor movement, or (2) with the cursor being returned to its original position for continued text production in cases when the cursor was moved. For example, if the writer types “colorless green ideas,” then moves the cursor elsewhere in the text (for example, to correct a typo), but then returns the cursor to the position after the word “ideas” and types “sleep furiously,” then the stretch of text “colorless green ideas sleep furiously” would be considered a single “block” of text. “Major blocks” were defined as blocks of text of at least 615 characters (or approximately 4 or more lines; see Appendix B)[2].

In line with the third improvement goal from Iteration 1, a semi-automated process was used to determine the remediation plans for participants in Iteration 2. Rather than manually inspecting the graphs displayed in Figures 2a–2d, the research team could now view 30 automatically extracted metrics, summarized in Appendix B. An example of all metrics for a session from one participant is shown in Appendix C. These metrics were extracted from the raw data using R scripts[3] based on a combination of keystroke logging (timings and result of the keystroke action), cursor movement, and eye tracking. The extracted metrics can be divided into four thematic groups: (1) how often did the participant revise their text, (2) did the participant display pausing and (3) reading / lookback behavior, and (4) to what extent participants engaged in non-linear writing. A matrix with the partial correlations of all measures separated into their thematic groups can be found in Appendix D. Metrics were calculated excluding the final revision, either across the entire writing process or individually for events that appear between sentences, between words, or before finishing a word. Participants’ writing-process measures were compared to a separately collected reference sample of 30 writing-process recordings (academic peers from the same population performing a similar task). We are agnostic as to the text quality in this reference sample because, unlike some of the previous studies (Baaijen and Galbraith 2018; Dux Speltz and Chukharev-Hudilainen 2021), our approach does not assume that there is a pattern of writing processes that is necessarily beneficial for all writers in all contexts. Instead, we treated deviations in the writer’s measures from the population norms (estimated using this reference sample) as possible remediation targets, but only if such deviations co-ocurred with (and plausibly caused) written-product deficits. In future work, a more representative sample will be obtained.

The semi-automated process analysis is illustrated by examining the data from one participant: Figure 6 shows the first six metrics as they are displayed on the ProWrite system for Participant 5’s diagnostic session. Rather than manually and qualitatively analyzing the writing-process playback as in Iteration 1, we examined point-range plots as in Figure 6 to determine how the participant compared to other participants in the reference sample in terms of specific writing-process behaviors. The shaded boxes represent the distribution of the reference sample for the metric, the blue dot represents the participant mean, and the range reflects predicted possible values for the participant (if the participant were to produce new texts in a similar writing context). In this case, Figure 6 shows that Participant 5 edited her text less often than other participants between sentences (Metric 1), between words (Metric 2), before finishing a word (Metric 3), and in general (Metric 4). She also produced more text between edits (Metric 5). By examining these metrics, we could

[2] The authors acknowledge Jennifer Godbersen who substantially contributed to the operationalization of linearity of writing for this study. For other operationalizations of linearity, see e.g. Hall, Baaijen and Galbraith (2022). Comparing different approaches to measuring linearity is an important topic that is, however, outside the scope of the present paper.

[3] See https://osf.io/x9b42/ for source code and sample data.

determine that this participant would not be a good candidate for the “Do not edit” remediation plan since she does not edit more than an average participant. Further examination of Participant 5’s writing-process metrics, shown in Figure 7, revealed that she produced fewer blocks of text (Metric 21) and fewer major blocks (Metric 22).

The metrics show that this participant rarely moved the cursor away from the leading edge to produce text or make revisions. Therefore, by inspecting these metrics, we determined that Participant 5 was a good candidate for one of the new remediation plans, “Revise periodically.” If a participant had more writing blocks and major writing blocks compared to the reference sample, this would indicate that they had a nonlinear writing process including many movements of the cursor away from the leading edge. Therefore, they would be a good candidate for the other new remediation plan, “Write linearly.” We have not yet developed a rule-based procedure for identifying specific metrics that would lead to the decision to assign a remediation plan. Therefore, we compared the participant to the reference sample and made an informed decision about an appropriate remediation plan. Further discussion of the metrics that led to decisions for specific remediation plans is included in the findings section below.

3.3.1.4 The Intervention Session

After we determined a remediation plan based on the participant’s diagnostic session, the participant returned for another 1-hour session to participate in an intervention. During this intervention session, a slightly revised protocol from Iteration 1 was used to present the new writing-process metrics to participants on a dashboard prototype. As in Iteration 1, the researcher began the session by allowing the participant to view the text that they had written in their diagnostic session. Then the researcher explained one or two areas of concern identified by the manual product analysis and explained that these issues may have emerged from issues that occurred during the writing process. The writing-process metrics were then presented to the participant, and the metrics most relevant to the selection of the remediation plan were introduced and explained. To illustrate the participant’s writing process, the researcher showed the participant brief replays of their writing process which included examples of the problematic behavior(s). Next, the researcher followed the same protocol as in Iteration 1, introducing the proposed remediation plan and explaining how the real-time scaffolded feedback would appear on the ProWrite system.

Mechanisms for triggering automated scaffolding for the two new remediation plans were added to the ProWrite system. The text of the feedback message for the “Do not edit” remediation plan was changed to “Commit to finishing your sentence.” The pop-up boxes for this iteration appeared as shown in Figures 8a–8d.

Upon the completion of the essay, the participant responded to a semi-structured interview pertaining to their experience with the ProWrite system. Participants were asked to what extent they perceived the intervention as useful, how receiving the automated feedback made them feel, how they responded to the feedback upon receiving it, how distracting they felt the feedback was, and whether they would consider using this strategy in future writing (and if so, in which stage of the writing process). The interview was audio-recorded, transcribed, and analyzed descriptively to determine patterns in participant responses.

3.3.1.5 The Follow-up Session

Finally, there was another important difference in the follow-up sessions of Iteration 2. Rather than saying that the participants could write using any strategy that they prefer, the researcher began the session by asking the participant whether they remembered the remediation plan from the intervention session. After the participant confirmed that they remembered the remediation plan, the researcher explained, "This time, you will not receive any feedback and we will not monitor your process, but we believe your text will improve if you use the strategy that you learned during our last session. Once again, the goal is to produce a good text." The goal of this instruction was to clarify that the researcher believed that the intervention strategy is beneficial, but the participant was given agency in deciding whether they would continue to use it. Upon the completion of the writing task, participants completed one more semi-structured interview. They were asked whether they intended to use the intervention strategy and whether they felt they were successful in implementing it.

3.3.2 Findings

Iteration 2 was guided by four research questions that emerged in response to the improvement plan developed upon the completion of Iteration 1. Findings will be presented below in light of each of the four research questions.

3.3.2.1 RQ1 findings

The first research question concerned how the automatic writing process metrics could be used to select a remediation plan. We developed a process to automatically extract metrics that allowed us to choose a remediation plan more quickly and less subjectively. It was no longer necessary to watch a keystroke-by-keystroke playback of the writing process and follow the aforementioned manual review process to select a remediation plan. Instead, the selection of the remediation plan was informed by point-range plots that show precisely how a participant's writing process compared to a reference sample of peers from the same population.

In Iteration 2, the process of determining the appropriate remediation plan was not fully automated, however. We still had to examine the point-range plots in order to determine the ways in which a participant differed from the reference sample. In future iterations, this process could be further automated by developing a rule-based mechanism for assigning remediation plans based on a synthesis of process measures.

Table 2 summarizes the metrics that were found to be most useful for illustrating writing-process behavior that stood out as potentially problematic and therefore to assign remediation plans. Other metrics were relevant for ruling out remediation plans that would not be appropriate for a participant.

3.3.2.2 RQ2 findings

The semi-automated extraction of writing-process metrics made it possible to measure the extent to which a participant adhered to the remediation plan during the intervention and follow-up sessions. To determine how participants' processes were impacted by the intervention, we compared each participant's writing-process metrics for all three writing sessions. Figure 9 illustrates the plot that was created for the writing blocks metric. Data from six participants are shown in Figure 9: three participants received the "Revise periodically" remediation plan; three received the "Write linearly" remediation plan.

The writing-process metric demonstrates that participants' writing processes were impacted by the remediation plans during the intervention stage. For example, as Figure 9 illustrates, participants who received the "Revise periodically" remediation plan initially showed fewer writing blocks than the reference sample (indicated by the dotted line) in the diagnostic session, whereas participants who received the "Write linearly" remediation plan displayed far more writing blocks than the reference sample in their diagnostic session. During the intervention session, participants that were assigned the "Revise periodically" remediation plan show an increased number of writing blocks. Participants that received the "Write linearly" remediation plan showed a substantially decreased number of writing blocks. Participants differed in terms of whether they sustained the writing-process strategy in the follow-up session. Most participants were closer to the reference sample in the intervention and follow-up sessions than in the diagnostic sessions, but they generally did not sustain the taught process behavior in the follow-up sessions to the same extent as in the intervention sessions. Similar graphs were created for all other metrics and showed similar tendencies but were omitted here because of space constraints.

To determine how participants' written products were impacted by the invention, we compared the product scores determined from the rubric described above. Figure 10 presents the scores for all eight participants for each session type and rating criterion. Participants' scores did not differ substantially across session types for most of the rating criteria. Participants 12 and 13 showed the most noticeable changes in their ratings across tasks. Generally, Participant 12 showed lower text quality scores at the intervention session compared to both the diagnostic session and follow-up session. Participant 13's introduction, prompt, and focus scores showed the same pattern, but her transition score dropped after the diagnostic session, and her conclusion score dropped after the intervention session. These findings seem to align with what these participants expressed in their post-intervention interviews, which we summarize in the next section.

In sum, this iteration showed that it is possible to modify participants' process behavior in response to remediation plans during and beyond an intervention session. However, there was no evidence that changing writing processes impacted text quality.

3.3.2.3 RQ3 findings

RQ3 concerned participants' experience with the new version of the ProWrite system. As in Iteration 1, all participants considered the intervention to be moderately useful (n=4) or very useful (n=4). Participant 12 rated the intervention as moderately useful ("2 or 3" out of 5). This participant was assigned the "Write linearly" remediation plan. She expressed her reason for considering it frustrating as "I consider myself a pretty good writer, [so] it was difficult to kind of have to restructure one of the primary ways that I write, which is to just throw all of the information I know at the page and then move it around so that it comes into order." However, she also expressed that the remediation plan was helpful for "forcing" her to put ideas more quickly into writing, which was important because there was a time constraint on the writing session. She also noted that the remediation plan "would definitely be something that's really useful for a rough draft stage."

Another participant rated the intervention as extremely useful (5 out of 5) because the "Revise periodically" remediation plan felt like a "natural" way for her to modify and improve her writing process. She also noted that when writing other timed essays (including in her diagnostic essay), she was aware that she prioritized putting ideas down in a "stream of consciousness" style and did not allow herself time to make revisions; she knew this was a problem for her. Adhering to the remediation plan improved her confidence in the quality of her text.

Seven out of the eight participants responded that they would try to use the newly learned strategies in their future writing experiences. However, they varied in terms of how they imagined themselves using the strategy in the future. For example, Participant 13 said that she would use the "Do not edit" strategy during the drafting process (i.e., the phase in which a writer is focused on generating content and ideas without prioritizing sentence-level correctness or style) because she felt that "not editing during the generative part will mean that when you finally do go back and edit, it will be better, and the things that you create during the generative process will be better." On the other hand, Participant 9 noted that she would use the "Pause sentence-initially" strategy in order to intentionally focus on her writing process while writing introductions and conclusions: "I would use this strategy more in the introduction [or] conclusion sections because that's where I really want to focus on what I'm going to be writing and what I'm trying to wrap up and summarize. When I'm drafting the body, I probably wouldn't use it as much just because I like to get it all out there and then work with what I have."

Six participants indicated that the pop-up boxes did not make it difficult to produce text. Specifically, participants indicated the pop-up boxes were not very distracting or even "distracting in a good way." Participant 11 noted that this automated feedback "did what it was intended to do, reminding me to go back," and Participant 13 highlighted that she noticed the prompt, but "it wasn't jolting me out of the writing process entirely." Participant 9 expressed that the pop-ups were very distracting to her, but only "for a brief period" until they were "easy to dismiss […] and get back to my writing."

In sum, all participants except one considered the intervention in Iteration 2 useful; one participant, however, recognized the potential of the suggested strategy for certain stages of the writing process. Most participants expressed that they would like to try to incorporate the strategy they learned in their future writing outside of the ProWrite system. Participants also expressed that the automated scaffolding was noticeable and helpful for reminding them of the remediation plan and the pop-up boxes did not disturb their ability to produce text.

3.3.2.4 RQ4 findings

RQ4 addressed whether returning participants sustained previously learned process modifications. Three participants—Participants 5, 6, and 7—participated in both Iteration 1 and Iteration 2, with 5–7 months in between iterations. Table 3 summarizes the remediation plans for these participants in both iterations and presents findings in response to RQ4.

All three participants showed improvement in terms of the process behavior identified as problematic during their Iteration 1 diagnostic session. Participants 5 and 6 were assigned the "Do not edit" remediation plan in Iteration 1 and showed substantial improvement in their editing behavior in Iteration 2. During Iteration 1, both participants edited far more than the average participant, but their editing behavior was consistent with the reference sample during their diagnostic session in Iteration 2. Participant 7 was assigned "Pause sentence-initially" in Iteration 1; during theIteration 2 diagnostic session, she maintained the remediation plan partially by pausing between sentences more frequently compared to the reference sample, but she reverted to her former process behavior and paused between mid-sentence words more often than average. In sum, these findings highlight the potential of these remediation plans to be sustained past the duration of the intervention session, potentially beyond automated scaffolding.

3.3.3 Reflection

The primary goals of Iteration 2 were as follows: 1) to determine how automatic writing-process metrics can be used for the selection of a remediation plan, 2) to analyze whether participants' writing processes and written products change based on the remediation plans, 3) to qualitatively assess how participants experience the new version of the ProWrite system, and 4) to determine whether returning participants sustain previously taught process modifications.

The semi-automatically extracted writing-process metrics allowed the research team to effectively and efficiently make decisions about remediation plans appropriate for participants based on behavioral data from their diagnostic session. They also made it possible to make between-participant and within-participant comparisons. In future iterations, it would now be possible to formulate a series of rules to assign remediation plans automatically.

Participants' writing processes were found to change in line with all four remediation plans. The written products did not improve substantially in the intervention session. Additionally, because most participants earned scores of 3 or 4 on the rubric, the rubric could not successfully discriminate texts of varying qualities. Therefore, this iteration demonstrated that the rubric should be modified for future iterations to capture differences in participants' skill levels.

All participants considered the intervention to be useful, and most participants (n=7) expressed that they would try to implement their remediation plan into their own writing. Participants described their remediation plans using terms such as "natural" and "helpful," and a few participants noted that they believed the plans helped them to produce better writing. These findings indicate that the system was generally positively perceived, and participants were not discouraged by the writing challenge. In fact, participants seemed to be motivated and intrigued by the opportunity to modify and potentially improve their writing processes using the assigned remediation plan.

Finally, Iteration 2 demonstrated that it is possible for returning participants to sustain previously learned process modifications. The three participants who returned from Iteration 1 to Iteration 2 exhibited writing-process behavior that demonstrated that they maintained at least some of the previously taught remediation plans. This occurred even though participants returned five to seven months after initial participation.

4 Discussion and Conclusion

This paper presented two design iterations of the ProWrite system which aims to combine concurrent keystroke logging and eye tracking data to generate individualized process-focused feedback to writers. This feedback was grounded in the analysis of each writer's individual needs and was presented in the context of a learning cycle consisting of an initial diagnostic assignment, an intervention assignment with an assigned remediation plan, and a follow-up assignment. Four remediation plans were developed and implemented in this study: "Do not edit," "Pause sentence-initially," "Revise periodically," and "Write linearly."

Several limitations were present in this study, presenting opportunities for future research. First, this study was limited in the number and kinds of remediation plans that were provided for participants and the feedback associated with the selected remediation plans. Therefore, we do not claim that the study investigates writing-process modifications in principle; instead, a limited set of remediation plans was evaluated. While we found that these remediation plans were effective for modifying participants' process behaviors with automated real-time scaffolding, our study did not aim to evaluate whether all aspects of the writing process can be successfully modified in the same way. Future work will continue to expand on the number and types of writing-process feedback to address this limitation.

Secondly, the current prototype of the ProWrite system is not fully automatic, and therefore is not immediately deployable at scale as it still requires a human analyst to review the process metrics, determine an appropriate remediation plan, and deliver that remediation plan to the participant. Future work will focus on automating the full analysis and intervention pipeline. This requires a better formalization of the rules that determine patterns of behavior that necessitate the assignment of a remediation plan. Furthermore, current automated process analysis is dependent upon statistical comparisons to a reference sample, with remediation plans being suggested when a participant deviates substantially from the estimated population norms. While our current approach is agnostic to the quality of texts in the reference sample, future work will investigate whether using high-quality texts might improve system performance. In Vandermeulen et al. (2020), for example, participants were shown how their writing processes differed from two benchmark writing processes (collected from a large national baseline study), one of which scored one standard deviation above the participant in text quality and the other scoring two standard deviations above the participant. It is possible that this approach of comparing processes to more advanced writers could benefit future iterations of the ProWrite system as well.

Finally, the current study lacked both the procedure and the sample size necessary to establish whether the remediation plans resulted in any differences in text quality. Our tentative approach of using an analytic rubric did not appear to be effective at detecting differences between diagnostic, intervention, and follow-up sessions in Iteration 2. Future consultation with assessment and writing pedagogy experts will be necessary to determine a better way to evaluate the efficacy of the intervention. It may be possible that the intervention, in its present form, does not lead to changes in text quality. This could be due to the fact that participants only received the automated scaffolding and process feedback once, making it difficult to lead to a lasting effect on follow-up writing. It may also be due to the fact that the system targeted only one writing-process behavior at a time. Galbraith and Baaijen (2018) suggest that text quality is predicted by a combination of interacting processes, and they recommend developing complex interventions that target multiple writing-process behaviors. Future research should therefore investigate whether combining remediation plans leads to better text quality. The next design iteration will prioritize improving the pedagogical components of the system and evaluating the system's potential for written-product improvement.

Despite its limitations, this study was the first one (to our knowledge) to demonstrate that concurrent keystroke logging and eye tracking can be used for delivering individualized remediation plans to learners, and such plans can be effectively scaffolded through real-time, automated prompting. Such real-time feedback can be useful because it allows for behavioral modification to occur while writing takes place instead of afterward, when it is arguably too late (i.e., after a student has already received a poor grade). This opportunity is answering the calls of previous studies to implement more immediate feedback to writers (Chukharev-Hudilainen et al. 2019; Conijn et al. 2020; Dux Speltz and Chukharev-Hudilainen 2021; Révész, Michel, and Lee 2019). However, since several participants did not maintain their writing-process changes during the follow-up session, future development of the system should also consider how to continue to encourage enacting remediation plans after the removal of automated scaffolding.

Future development of the ProWrite system should also consider the feedback from Iteration 2 participants who noted that their remediation plans could be more useful for them during specific parts of their writing process. For example, a few participants noted that their remediation plan could help them in their future writing during the "rough drafting" or "generative writing" stages, and another participant considered applying the remediation plan while writing introductions and conclusions. This could also alleviate some participants' frustrations about completely abandoning successful parts of their current writing processes.

Overall, the ProWrite system and its process-focused feedback was viewed overwhelmingly positively by participants. Future development of the ProWrite system will prioritize expanding and refining diagnostic rules for remediation plan selection in order to further automate this process and improve the system's efficacy for improving text quality. This development has the potential to make the ProWrite system scalable and effective for classroom applications, allowing students to receive individualized writing feedback with relatively little involvement of teachers. Furthermore, after several additional iterations in which we will continue to automate and expand upon ProWrite's capabilities, a summative evaluation will be conducted to determine whether the individualized, process-focused feedback provided by ProWrite improves text quality, over and above benefits that are afforded by individualized product-focused writing coaching (i.e., the current practice of university writing centers).

5 References

Anderson, T., and Shattuck, J. (2012). Design-Based Research: A Decade of Progress in Education Research? *Educational Researcher* 41 (1): 16–25. https://doi.org/10.3102/0013189X11428813.

Baaijen, V. M., Galbraith, D., and De Glopper, K. (2012). Keystroke Analysis: Reflections on Procedures and Measures. *Written Communication* 29 (3): 246–277.

Baaijen, V. M., & Galbraith, D. (2018). Discovery through Writing: Relationships with Writing Processes and Text Quality. *Cognition and Instruction* 36 (3): 199–223.

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). https://doi.org/10.1111/modl.12316

———. 2019. What Can L2 Writers' Pausing Behavior Tell Us about Their L2 Writing Processes? *Studies in Second Language Acquisition* 41 (3): 529–54. https://doi.org/10.1017/S027226311900010X.

Bennett, R. E., Zhang, M., Deane, P., & van Rijn, P. W. (2020). How do Proficient and Less Proficient Students Differ in their Composition Processes? *Educational Assessment* 25 (3): 198–217. https://doi.org/10.1080/10627197.2020.1804351.

Bowen, N., Thomas, N., and Vandermeulen, N. (2022). Exploring Feedback and Regulation in Online Writing Classes with Keystroke Logging. *Computers and Composition* 63 (March): 1–30. https://doi.org/10.1016/j.compcom.2022.102692.

Bowen, N., and Van Waes, L. (2020). Exploring Revisions in Academic Text: Closing the Gap Between Process and Product Approaches in Digital Writing. *Written Communication* 37 (3): 322–64. https://doi.org/10.1177/0741088320916508.

Breetvelt, I., van den Bergh, H., and Rijlaarsdam, G. (1994). Relations between Writing Processes and Text Quality: When and How? *Cognition and Instruction* 12 (2): 103–23. https://doi.org/10.1207/s1532690xci1202_2.

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *Journal of the Learning Sciences* 2 (2): 141–78. https://doi.org/10.1207/s15327809jls0202_2.

Chukharev-Hudilainen, E., and Saricaoglu, A. (2016). Causal Discourse Analyzer: Improving Automated Feedback on Academic ESL Writing. *Computer Assisted Language Learning* 29 (3): 494–516. https://doi.org/10.1080/09588221.2014.991795.

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., and Feng, H-H. (2019). "Combined Deployable Keystroke Logging and Eyetracking for Investigating L2 Writing Fluency." *Studies in Second Language Acquisition* 41 (3): 583–604. https://doi.org/10.1017/S027226311900007X.

Collins, A. (1992). Toward a Design Science of Education. In *New Directions in Educational Technology*, 15–22. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-77750-9_2.

Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van Waes, L., and van Zaanen, M. (2020). How to Provide Automated Feedback on the Writing Process? A Participatory Approach to Design Writing Analytics Tools. *Computer Assisted Language Learning*, November, 1–31. https://doi.org/10.1080/09588221.2020.1839503.

Duncheon, J. C., and Tierney, W. G. (2014). Examining College Writing Readiness. *The Educational Forum* 78 (3): 210–30. https://doi.org/10.1080/00131725.2014.912712.

Dux Speltz, E., and Chukharev-Hudilainen, E. (2021). The Effect of Automated Fluency-Focused Feedback on Text Production. *Journal of Writing Research* 13 (2): 231–55. https://doi.org/10.17239/jowr-2021.13.02.02.

Feng, H-H. (2015). Designing, Implementing, and Evaluating an Automated Writing Evaluation Tool for Improving EFL Graduate Students' Abstract Writing: A Case in Taiwan. Edited by Carol A. Chapelle and Evgeny Chukharev-Hudilainen. Ann Arbor, United States: Iowa State University. https://www.proquest.com/dissertations-theses/designing-implementing-evaluating-automated/docview/1762721956/se-2.

Feng, H.-H., Chukharev-Hudilainen, E. (2017, September). Tailoring writing pedagogy in light of ESL students' pausing behavior during the writing process. Paper presented at the 15th Technology for Second Language Learning Conference at Iowa State University, Ames, Iowa.

Galbraith, D., and Baaijen, V. M. (2019). Aligning Keystrokes with Cognitive Processes in Writing. In *Observing Writing*, 306–25. Brill. https://doi.org/10.1163/9789004392526_015.

Graham, S., and Harris, K. R. (2003). Students with Learning Disabilities and the Process of Writing: A Meta-Analysis of SRSD Studies. In *Handbook of Learning Disabilities , (pp*, edited by H. Lee Swanson, 587:323–44. New York, NY, US: The Guilford Press, xvii.

Graham, S., McKeown, D., Kiuhara, S., and Harris, K. R. (2012). A Meta-Analysis of Writing Instruction for Students in the Elementary Grades. *Journal of Educational Psychology* 104 (4): 879–96. https://doi.org/10.1037/a0029185.

Graham, S., and Perin, D. (2007). A Meta-Analysis of Writing Instruction for Adolescent Students. *Journal of Educational Psychology* 99 (3): 445–76. https://doi.org/10.1037/0022-0663.99.3.445.

Hall, S., Baaijen, V. M., and Galbraith, D. (2022). Constructing theoretically informed measures of pause duration in experimentally manipulated writing. *Reading and Writing* 2022: 1-29.

Hayes, J. R., and Flower, L. S. (1980). Identifying the Organization of Writing Processes. In *Cognitive Processes in Writing*, edited by L. W. Gregg and E. R. Steinberg, 3–30. Hillsdale, NJ: Lawrence Erlbaum.

Leijten, M., and Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication* 30 (3): 358–92. https://doi.org/10.1177/0741088313491692.

Ranalli, J., Feng, H.-H., and Chukharev-Hudilainen, E. (2018). Exploring the Potential of Process-Tracing Technologies to Support Assessment for Learning of L2 Writing. *Assessing Writing* 36 (April): 77–89. https://doi.org/10.1016/j.asw.2018.03.007.

———. 2019. The Affordances of Process-Tracing Technologies for Supporting L2 Writing Instruction. *Language Learning & Technology* 23 (2): 1–11.

Reeves, T. C., and McKenney, S. (2013). Computer-Assisted Language Learning and Design-Based Research: Increased Complexity for Sure, Enhanced Impact Perhaps. In *Design-Based Research in CALL*, 9–21. CALICO, The Computer Assisted Language Instruction Consortium. https://research.utwente.nl/en/publications/computer-assisted-language-learning-and-design-based-research-inc.

Révész, A., Michel, M., and Lee, M. (2019). Exploring Second Language Writers' Pausing and Revision Behaviors: A Mixed-Methods Study. *Studies in Second Language Acquisition* 41 (3): 605–31. https://doi.org/10.1017/S027226311900024X.

Rijlaarsdam, G., and Van den Bergh, H. (2006). Writing Process Theory. *Handbook of Writing Research*, 41–53.

Rogers, L. A., and Graham, S. (2008). A Meta-Analysis of Single Subject Design Writing Intervention Research. *Journal of Educational Psychology* 100 (4): 879–906. https://doi.org/10.1037/0022-0663.100.4.879.

Torrance, M. (2015). Understanding Planning in Text Production. *Handbook of Writing Research*, 1682–90.

Vandermeulen, N. (2020). Synthesis Writing in Upper-Secondary Education: From a Baseline of Texts and Processes to Process-Oriented Feedback. Antwerp, Belgium: University of Antwerp. https://lirias.kuleuven.be/3070762?limo=0.

Vandermeulen, N., Leijten, M., and Van Waes, L. (2020). Reporting Writing Process Feedback in the Classroom: Using Keystroke Logging Data to Reflect on Writing Processes. *Journal of Writing Research* 12 (1), 109–40. https://doi.org/10.17239/jowr-2020.12.01.05

Zhang, M., Guo, H., Liu, X., and MZhang, H. (2017). Using Keystroke Analytics to Understand Cognitive Processes during Writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1568–78.

6 List of Figures

7 Tables

Table 1. Features contributing to each remediation plan's assignment.

	"Do not edit"	"Pause sentence-initially"
Final review	Often absent or limited	Varied
Episodes of inscription	Frequent	Varied
Time that writing begins	Varied	Often immediate
Time between first read-through of the prompt and the beginning of writing	Varied	Often immediate
Frequency/volume of deletions	Frequent and/or large deletions	Varied
Location of longest pauses	Between sentences or clauses	Mid-sentence or mid-word
Sentence-initial pauses	Present	Absent

Table 2. Remediation plans from Iteration 2 and the metrics that determine their assignment.

Remediation Plan	**Relevant Metrics**
Do not edit	How often did you edit your text before finishing a word? How often did you edit your text?
Pause sentence-initially	How long were your between-sentence pauses? How often did you pause between words?
Write linearly	How many times did you jump between text chunks? How many different writing blocks did you create? How often did you edit your text between words?
Revise periodically	How often did you look back into your text? How much text did you produce without looking back into your text? How many different writing blocks did you create? How many major writing blocks did you produce?

Table 3. Overview of participants who participated in both iterations.

	Participant 5	**Participant 6**	**Participant 7**

Iteration 1 remediation plan	Do not edit	Do not edit	Pause sentence-initially
Iteration 2 remediation plan	Revise periodically	Write linearly	Revise periodically
Did they sustain the Iteration 1 remediation plan in Iteration 2?	Yes. Process analysis for her Iteration 2 diagnostic session revealed less frequent editing than the average participant.	Yes. Process analysis for her Iteration 2 diagnostic session revealed editing behavior consistent with the average participant.	Somewhat. Process analysis for her Iteration 2 diagnostic session revealed that she paused between sentences more frequently than the average participant, but she also had very short between-sentence pauses and paused between words more often than average.

8 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

9 Funding

This material is based upon work supported by the National Science Foundation under Grant No. 2016868. The open access publication fees for this article were covered by the Iowa State University Library.