



On the role of engagement in automated feedback effectiveness: Insights from keystroke logging

Ronja Schiller^{a,*}, Johanna Fleckenstein^b, Lars Höft^a, Andrea Horbach^a, Jennifer Meyer^{a,c}

^a Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118, Kiel, Germany

^b University of Hildesheim, Department of Applied Educational Science, Universitätsplatz 1, 31141, Hildesheim, Germany

^c University of Vienna, Department for Teacher Education, Porzellangasse 4, 1090, Vienna, Austria

ARTICLE INFO

Keywords:

Automated feedback
Behavioral and cognitive learning engagement
Process-oriented
Text revision
Keystroke logging

ABSTRACT

Feedback research increasingly focuses on the role of learners' engagement in the feedback process. Process measures from technology-based learning environments that reflect writing behavior can provide new insights into the mechanisms underlying feedback effectiveness by making engagement visible. Previous research has shown that log data and similarity measures mediate the effects of automated feedback on learners' revision performance. In the present study, we aimed to replicate and extend previous research using measures obtained from keystroke logging that represent the revision process on a more fine-grained level. We considered behavioral engagement (i.e., number of keystrokes and typing time) and writing pauses as potential indicators of cognitive engagement. In a classroom experiment, $N = 453$ English-as-a-foreign-language (EFL) learners ($M_{\text{age}} = 16.11$) completed a writing task and revised their draft, receiving either feedback generated by a large language model (i.e., GPT 3.5 Turbo) or no feedback. A second writing task served as a transfer task. All texts were scored automatically to assess performance. The effect of automated feedback on learners' revision and transfer performance was mediated through the different indicators of behavioral engagement during the text revision, although the direct effect of automated feedback on the transfer task was not significant. We found small effects of feedback on pause length and the number of pauses, but the indirect effects were not significant. The study provides further evidence on the role of learning engagement in feedback effectiveness and illustrates how online measures (i.e., keystroke logging) can be used to gain new insights into the effectiveness of automated feedback. The use of different process measures to assess learning engagement is discussed.

1. Introduction

Automated feedback is a powerful tool for improving learning in various domains (Mertens et al., 2022; Van der Kleij et al., 2015). Also in writing, there is ample evidence that learners benefit from automated feedback in improving their performance (Fleckenstein et al., 2023; Ngo et al., 2024; Zhai & Ma, 2022). Feedback research has mainly focused on performance-related outcomes and we do not know exactly how feedback works (Winstone & Nash, 2023). Recently, feedback research has begun focusing more on the active role of

* Corresponding author. Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118, Kiel, Germany.
E-mail address: schiller@leibniz-ipn.de (R. Schiller).

the learner, which is considered to be crucial to understand what makes feedback effective (Panadero, 2023; Price et al., 2011; Winstone et al., 2017), highlighting the role of feedback in increasing learning engagement on the task level. Technology-based learning environments hold great potential for assessing learning engagement, thereby facilitating a deeper understanding of learning processes (Henrie et al., 2015). The domain of writing makes it possible to collect particularly fine-grained data that reflect learners' behavior during a task (Leijten & Van Waes, 2013). Therefore, conducting computer-based studies related to writing is a great opportunity to follow calls from feedback research to investigate the effectiveness of feedback from a process-oriented perspective and to gain more in-depth insights into the role of learning engagement (Price et al., 2011; Reinhold et al., 2024; Winstone et al., 2017; Winstone & Nash, 2023).

Previous research has already provided first evidence that process data based on log data (i.e., the total time on task) and text similarity measures (i.e., the edit distance) can be useful to uncover what learners do during text revision if they receive automated feedback compared to a control group (Schiller et al., 2024). Here, the total time on task and the edit distance were used as indicators of the behavioral dimension of learning engagement; however, these measures are somewhat limited in their informative value. During the total time on task, it is unclear whether a student is actively working or just has a particular page open in a survey tool (Schiller et al., 2024). Furthermore, the edit distance is calculated on the basis of two texts after they have been written (Levenshtein, 1965), which is why information such as repeated deletions of characters is not included in the edit distance. Deletion and substitution are important parts of text revision; therefore, referring to measures such as the edit distance could lead to the underestimation of learners' behavioral engagement (Faigley & Witte, 1981; Lindgren & Sullivan, 2006). Measures that not only count revisions that are visible in the final text but also consider repeated deletions or substitutions of characters would capture learners' behavioral engagement during the text revision in a more complete way.

The method of keystroke logging provides additional opportunities to obtain process measures that reflect writing behavior more comprehensively (Conijn et al., 2021; Leijten & Van Waes, 2013). Keystroke logging data contain information about the position in the text and the timepoint of each keystroke made by a learner during the writing process, providing detailed behavioral data of the writing process (Leijten & Van Waes, 2013). With this information, it is possible to retrieve, for example, the typing time, in addition to the total time on task. Moreover, keystroke logging considers editing behavior, which does not become visible when referring, for example, to the edit distance (Levenshtein, 1965). For instance, keystroke logging provides information about the extent to which learners deleted or substituted text (Conijn et al., 2021), while the edit distance only includes those changes to the text that can be derived from the end products of two text versions.

Previous studies have focused primarily on the behavioral dimension of learning engagement, partly because its definition via observable behavior makes it straightforward to operationalize (Wong et al., 2024). However, it is important to also consider other dimensions of learning engagement than just behavioral engagement to better understand what learners do during a task (Wong et al., 2024). In the area of writing, inactive periods (i.e., writing pauses) are often interpreted as phases of cognitive processes such as planning and reflection that are central to the writing process as well (Flower & Hayes, 1981; Lindgren & Sullivan, 2006). In the field of writing research, it is common to include pausing behavior in the analyses of, for example, different writing patterns (e.g., Guo et al., 2018; Talebinamvar & Zarrabi, 2022; Zhang et al., 2019). Previous findings suggest that writers often show increased pausing behavior during more demanding parts of writing tasks, which indicates that writing pauses are associated with mental effort (e.g., Lacruz et al., 2012; Medimorec & Risko, 2017; Mohsen & Qassem, 2020). Thus, analyzing writing pauses obtained from keystroke logging data could be a possibility to capture processes associated with cognitive learning engagement. It should be noted that writing pauses do not always reflect cognitive processes. They can also be associated with off-task behavior, and their meaning can differ depending on the context (de Smet et al., 2018; Lindgren & Sullivan, 2006; Révész et al., 2019).

In sum, keystroke logging might provide better opportunities to operationalize the psychological construct of learning engagement because it allows researchers to retrieve measures that reflect learners' actual behavior in a more precise and comprehensive way. However, it has rarely been used to investigate learning engagement in relation to automated instructional feedback.

In the following, we refer to the edit distance and the total time on task as *offline measures* to simplify the distinction between these measures and the measures that we obtained from keystroke logging files. The edit distance can only be collected after the revision process has been concluded. The total time on task is also derived from log files; however, as described above, these log files contain less information about learners' actual revision behavior than keystroke logging data do. Regarding the measures obtained from keystroke logging files, we refer to *online measures* in the following.

In the present study, we had two major aims: First, we aimed to replicate prior research to provide further evidence on the role that behavioral learning engagement, assessed by offline measures (e.g., edit distance), plays in the effectiveness of automated feedback. Replication is crucial in research to establish the reliability and generalizability of empirical findings, ensuring that observed effects are robust and reproducible across different learner populations and educational contexts (Open Science Collaboration, 2015). In the present study, for instance, the automated feedback was generated by a large language model (LLM), in contrast to previous research that has mostly investigated automated writing evaluation (AWE) based on machine learning algorithms (e.g., Bennett & Zhang, 2015; Crossley et al., 2022; Shermis, 2014; Wilson & Czik, 2016). Furthermore, different age groups of learners can differ in the extent to which they profit from automated feedback on writing (Fleckenstein et al., 2023), which is why replicating findings with samples of learners of different age groups and across different genres is essential in the field of educational psychological research. Second, aiming to extend prior research, we used keystroke logging data to investigate whether online measures, as more detailed process measures, can add to the understanding of learning engagement as an important mechanism of feedback effectiveness. That is, we used online measures of behavioral engagement (i.e., total number of keystrokes and typing time), as well as additional measures related to writing pauses (i.e., total pausing time, mean pause length, and number of pauses).

By replicating previous research with a different sample, different tasks, and different feedback, and extending prior research by

using additional process measures, we aimed to verify prior findings while also addressing the need for more empirical research on automated feedback effectiveness from a process-oriented perspective, focusing on learning engagement during text revision as a central mechanism in automated feedback effectiveness (Panadero, 2023; Winstone & Nash, 2023).

2. Research background

2.1. A process-oriented perspective on automated feedback

Feedback informs a learner about their current and the target state regarding a task and provides information on how to close the gap between these states (Hattie & Timperley, 2007). Recently, it has been discussed in the feedback literature that the provision of feedback alone does not lead to improved performance (Winstone et al., 2017); the feedback has to be processed by a learner (Lipnevich & Smith, 2022). Learners who work with feedback more actively may understand the feedback better and then apply the information to their draft (Lipnevich & Smith, 2022; Narciss, 2008; Panadero & Lipnevich, 2022). This, in turn, might be associated with learners engaging in a task to a higher extent. For instance, receiving feedback for a writing task is associated with higher engagement during the text revision (Schiller et al., 2024). The more a learner engages in a task, the higher the chance is that they will improve their performance during this task (Winstone et al., 2017). In line with the current developments in the field of feedback research (Panadero, 2023; Price et al., 2011; Winstone et al., 2017; Winstone & Nash, 2023), we adopted a process-oriented perspective on feedback, considering learning engagement as the central mechanism underlying the effectiveness of (automated) feedback.

Especially in writing tasks, feedback is an important means of supporting learners in developing and improving their skills (Graham et al., 2015). Automated feedback allows for the regular and immediate provision of effective feedback on writing assignments (e.g., Fleckenstein et al., 2023; Ngo et al., 2024; Zhai & Ma, 2022). LLMs can generate elaborate feedback on written texts within seconds and such feedback has been shown to support learners in improving their writing (e.g., Chan et al., 2024; Meyer et al., 2024). Accordingly, in the context of this paper, we understand automated feedback to refer to feedback that is generated by an LLM.

With regard to the empirical investigation of mechanisms such as learning engagement, the implementation of (feedback) studies in digital settings allows for the collection of different kinds of process data. Process data have been used extensively in the field of learning analytics to analyze learners' interactions with digital environments (e.g., Biedermann et al., 2023; Choi et al., 2023; Gašević et al., 2017; Kizilcec et al., 2017; Quick et al., 2020; Winter et al., 2024). Sailer et al. (2024) emphasized the importance of more interdisciplinary research to connect analyses of this kind of data with appropriate theoretical foundations. For instance, process data hold great potential to retrieve indicators of psychological learning mechanisms such as task-level engagement (Henrie et al., 2015).

2.2. Learning engagement

In broad terms, learning engagement means learners' participation in learning-related situations (Fredricks et al., 2004). These situations range from more global contexts such as the school community to more narrow contexts such as the classroom or extra-curricular working groups to very narrow contexts such as working on specific tasks (Wong & Liem, 2022). In this sense, Wong and Liem (Wong & Liem, 2022) differentiated between "learning engagement" and "school engagement" and emphasized that they cannot be used synonymously. School engagement, on the one hand, is related to the psychological sense of belonging in the school and to relationships with peers and teachers but also to involvement in the school community and adherence to the school's rules (Fredricks et al., 2004; Wong & Liem, 2022). Learning engagement, on the other hand, refers more directly to processes during specific learning activities (Wong & Liem, 2022), such as working on a specific task with feedback. Therefore, in the present article, which focuses on learners' behavior during a specific revision task, we refer to learning engagement.

Learning engagement is conceptualized as a three-dimensional construct, consisting of an affective, a cognitive, and a behavioral dimension (Fredricks et al., 2004; Wong et al., 2024). The affective dimension of engagement reflects emotional reactions to specific learning activities (Fredricks et al., 2004; Wong & Liem, 2022). The cognitive dimension refers to strategies used by the learner to approach a specific task (Han & Hyland, 2019). The behavioral dimension refers to the actual observable behavior the learner shows during a learning activity (Fredricks et al., 2004). Although this three-dimensional conceptualization is well established, many studies refer to engagement as a general concept (Henrie et al., 2015; Wong et al., 2024). This overgeneralization of engagement (Wong et al., 2024) is accompanied by an inconsistency in definitions of the dimensions of engagement and their operationalization; this can result in studies using the same measures to capture different facets of engagement (Wong et al., 2024).

To address this problem, Wong et al. (2024) as well as Wong and Liem (2022) recommend that researchers focus on specific dimensions of engagement instead of referring to learning engagement as a general term and that they thoroughly consider its operationalization. So far, its clear definition makes behavioral engagement the dimension with the greatest consensus regarding the operationalization (Wong et al., 2024). Wong et al. (2024) found that most previous research on learning engagement has focused on the behavioral dimension, although the dimension was not always specified in the corresponding studies. According to Wong et al.'s (2024) meta-analysis, behavioral engagement, as opposed to affective and cognitive engagement, is also most strongly associated with learners' performance. Behavioral engagement, thus, can be considered to be the ability of a learner to translate their intentions into actions, which is decisive for their performance (Martin, 2023; Miller, 2015). However, with regard to writing, in particular, periods of planning and reflection are just as important as proactive observable editing behavior (Flower & Hayes, 1981). Phases of planning or reflection are associated with the cognitive dimension of engagement (Fredricks et al., 2004). Often, cognitive processes during writing are not directly observable and are considered to occur during typing pauses (Lindgren & Sullivan, 2006). However, although pauses

are recognized as indicators of cognitive processes in writing research (e.g., Medimorec & Risko, 2017; Mohsen & Qassem, 2020; Zhang et al., 2019), they could also indicate off-task behavior, similar to other time-related measures (Kovanovic et al., 2016; Lindgren & Sullivan, 2006). At the same time, cognitive processes do not necessarily occur during writing pauses, but can also be reflected in observable writing behavior (i.e., keystroke patterns) associated with specific phases of the writing process such as planning or drafting and occur in parallel with proactive typing (Alves et al., 2008; Fan et al., 2023; Hall et al., 2024; Roeser et al., 2025; Xu, 2024). For example, fluent and disfluent typing or inter-key- and inter-word intervals of different lengths at certain locations in the text could be associated with cognitive processes at different levels, like reflecting on content versus language-related aspects (Hall et al., 2024; Roeser et al., 2025). With regard to the measurement of learning engagement in the context of writing, this illustrates the challenges associated with capturing the behavioral and the cognitive dimensions of learning engagement as distinct constructs (Wong et al., 2024). When studying learning engagement, it should be clearly distinguished between the different dimensions of learning engagement and their operationalization to avoid the under- or overestimation of the impact of specific facets of learning engagement (Wong et al., 2024).

Behavioral measures are based on data that reflect learners' actual behavior. For instance, the time on task has been used previously as an indicator of effort or learning engagement in different contexts (Bråten et al., 2022; Järvelä et al., 2008; Zhu et al., 2020). The total time on task alone is a vague indicator of behavioral engagement because it masks whether learners actually work on a task or whether they are distracted and it is particularly difficult to interpret in the context of technology-based learning (Kovanovic et al., 2016). As Kovanovic et al. (2016) pointed out, showing off-task behavior in digital settings does not mean that learners are not engaged in a learning activity. It is possible that they are thinking about a problem related to the learning activity or that they are looking something up in their analogous material (Kovanovic et al., 2016). Although the total time on task has been criticized for its vagueness as it includes time during which learners may have been distracted, it is still an informative measure as it also includes the time during which learners were engaged in thinking or planning (Kovanovic et al., 2016). Writing is a process and, in the context of this process, these activities of thinking and planning are just as important as the writing itself (Flower & Hayes, 1981).

In sum, the total time on a task is an important indicator of learning engagement, but it can be difficult to interpret. Therefore, when aiming to gain a more complete picture of learning engagement during text revision in relation to automated feedback, it is necessary to use additional measures that shed more light on what learners actively do (during the total time on task).

Further, in writing tasks in particular, the extent of text revision behavior can be used to measure behavioral learning engagement (Horbach et al., 2022; Nguyen et al., 2017) and to complement time on task measures. The comparison of initial drafts and revised drafts allows for the calculation of the edit distance (Levenshtein, 1965). The edit distance provides information about how many revisions a learner has made to a text during a text revision (Horbach et al., 2022; Lacruz et al., 2012; Nguyen et al., 2017). Nguyen et al. (2017) and Lacruz et al. (2012), for instance, used the number of revisions as a measure of the effort invested during text revision, and Horbach et al. (2022) referred to the edit distance in association with engagement during text revision. As explained above, we refer to the edit distance and the total time on task as offline measures to simplify the distinction between these measures and the measures that we obtained from keystroke logging files (online measures).

With regard to online measures, keystroke logging has been used as a method to study writing behavior for many years (Leijten & Van Waes, 2013; Van Waes et al., 2014). Keystroke logging files contain information about each key pressed during the writing process, as well as the position of the key in the text and a timestamp for the keystroke (Leijten & Van Waes, 2013). Based on this information, it is possible to obtain online measures that reflect behavioral processes during text revision more precisely than the offline measures mentioned above do. In contrast to the edit distance, the total number of keystrokes considers not only deletions, insertions, and substitutions of characters that become visible when comparing two versions of a text (Levenshtein, 1965) but also how often a learner pressed a key to delete or substitute a character (Leijten & Van Waes, 2013). Keystroke logging also allows for the differentiation between the total time on task (i.e., the time from reaching a specific page of an online survey until clicking the forward button) and, for example, the typing time (i.e., the time from the first keystroke until the last keystroke; Zhang et al., 2019). Furthermore, measures related to writing pauses can be retrieved from keystroke logging data. Writing pauses could provide information about the extent to which learners are engaged in, for example, planning or reflecting (Lacruz et al., 2012; Lindgren & Sullivan, 2006; Medimorec & Risko, 2017; Mohsen & Qassem, 2020), making them potential indicators of processes associated with the cognitive dimension of learning engagement.

2.3. Using process data to understand the role of learning engagement in the effectiveness of automated feedback on writing performance

Automated feedback based on AWE has been shown to be effective with regard to learner performance in several meta-analyses (Fleckenstein et al., 2023; Graham et al., 2015; Ngo et al., 2024; Zhai & Ma, 2022). With regard to the mechanisms underlying these effects, a few previous studies have used different process measures to investigate the relationship between learning engagement (or related constructs such as effort), automated feedback, and writing performance.

2.3.1. Offline measures of learning engagement during text revision

For example, Fleckenstein et al. (2024) referred to the time learners spent on a page with automated feedback as an indicator of behavioral engagement. Although there were no significant differences regarding the effectiveness of different feedback types, they found a significant indirect effect of behavioral engagement, suggesting that the time learners spent on the feedback page mediated the effectiveness of individualized feedback on learners' writing performance (Fleckenstein et al., 2024). Nguyen et al. (2017) used the edit distance as an indicator of the effort invested during text revision in connection with different types of feedback, but no associations were found between the edit distance and the text quality. Schiller et al., 2024 used the total time on task and the edit distance

to measure learning engagement during text revision in connection with automated feedback. In contrast to [Nguyen et al. \(2017\)](#), they found significant positive effects of automated feedback on both of these measures, whereby the positive effect of the automated feedback on the learners' revision performance was mediated by the level of learning engagement ([Schiller et al., 2024](#)). In a study by [Zhu et al. \(2020\)](#), the number of revisions learners made was positively associated with performance improvement. They also found that learners showed higher numbers of revisions if they received more specific automated feedback compared to generic feedback, but the different types of feedback were not associated with differences in the extent of the performance improvements ([Zhu et al., 2020](#)).

2.3.2. Online measures of learning engagement during text revision

Using keystroke logging, [Zhang et al. \(2019\)](#) analyzed how different patterns of writing behavior are associated with text quality. They found that a higher number of total keystrokes and a longer typing time were related to performance. Moreover, their analyses showed that fluency (i.e., typing speed) can also be decisive for text quality ([Zhang et al., 2019](#)). Among learners with the same fluency, they found two different clusters: in one of these clusters, learners showed a higher number of keystrokes and better text quality than learners in the cluster showing a lower number of keystrokes. On the basis of this finding, [Zhang et al. \(2019\)](#) emphasized the role that persistence—or behavioral engagement—in writing plays in learners' writing performance. Here, it was not decisive whether a learner was capable of writing fast (i.e., fluency) but whether they showed persistence during the writing process (i.e., a high number of keystrokes).

Using keystroke logging, it is also possible to analyze writing pauses in addition to proactive revision behavior (e.g., [Lacruz et al., 2012](#); [Medimorec & Risko, 2017](#); [Mohsen & Qassem, 2020](#); [Vandermeulen et al., 2024](#)). Writing pauses have been interpreted as phases of cognitive processes, such as reflecting and planning, or as cognitive effort ([Conijn et al., 2021](#); [Lindgren & Sullivan, 2006](#)). A study by [Lacruz et al. \(2012\)](#), for example, showed that writers paused more during the revision of demanding compared to simpler text sections, which could indicate that an increase in writing pauses reflects increased cognitive effort. A study by [Medimorec and Risko \(2017\)](#) suggested that learners who pause more during writing invest more cognitive effort into formulating complex sentences ([Medimorec & Risko, 2017](#)). [Mohsen and Qassem \(2020\)](#) found that writers who wrote argumentative essays spent more time on pauses than those who wrote descriptive texts, implying that more complex writing tasks require more phases of cognitive processing than less complex writing tasks do. Similarly, [Vandermeulen et al. \(2024\)](#) found that learners who wrote argumentative essays showed a longer total pausing time than learners who wrote narrative texts.

The investigation of writing pauses has mostly been related to writing first drafts of writing tasks (e.g., [Alves et al., 2008](#); [Guo et al., 2018](#); [Medimorec & Risko, 2017](#); [Zhang et al., 2019](#)), while there is not much research on pausing behavior related to text revisions in the context of automated feedback.

2.3.3. Analyzing feedback processes using keystroke logging

In sum, keystroke logging data offer various possibilities for inferring potential indicators of learning engagement at the task level. In relation to teacher feedback, for instance, [Bouwer and Dirkx \(2023\)](#) used keystroke logging to identify different strategies for processing feedback. On the basis of keystroke logging data, they found that learners' revision behavior differed with regard to the extent to which they integrated the feedback into their texts (i.e., superficial versus meaningful revisions; [Bouwer & Dirkx, 2023](#)). However, only few studies so far have used keystroke logging to investigate learners' behavior in relation to automated instructional feedback. The existing studies used keystroke logging to provide learners with feedback on their writing process but not to investigate the role of learning engagement in the feedback's effectiveness. For example, [Dux Speltz and Chukharev-Hudilainen \(2021\)](#) used keystroke logging to deliver immediate feedback, based on the detection of writing pauses, to increase learners' fluency (i.e., keystrokes per minute) during essay writing. In their study, receiving feedback meant that the screen faded if a writing pause was detected and that the text reappeared as soon as the student resumed typing ([Dux Speltz & Chukharev-Hudilainen, 2021](#)). [Dux Speltz and Chukharev-Hudilainen \(2021\)](#) found that the learners who received this type of feedback showed higher fluency than the learners in a control condition; however, the accuracy of their writing was lower, probably because they felt time pressure ([Dux Speltz & Chukharev-Hudilainen, 2021](#)). [Vandermeulen et al. \(2020, 2023\)](#) also used keystroke logging as a basis for providing automated feedback on learners' writing processes and found that learners who received feedback on their writing process (e.g., regarding the time used for text production or for switching between sources) performed better in subsequent writing tasks ([Vandermeulen et al., 2020, 2023](#)). However, to the best of our knowledge, no studies have used online measures to capture learning engagement in response to instructional task-specific automated feedback, as it is conceptualized in the feedback literature ([Hattie & Timperley, 2007](#)).

We argue that using keystroke logging to measure learning engagement during text revision can extend prior research that used offline measures to investigate behavioral learning engagement as a key mechanism in the effectiveness of automated instructional feedback. Beyond that, we think that using keystroke logging to analyze writing pauses would contribute to our understanding of the cognitive dimension of learning engagement and the meaning of writing pauses in the context of revising and receiving feedback.

Considering the finding that more cognitively demanding writing tasks are associated with increased pauses ([Lacruz et al., 2012](#); [Medimorec & Risko, 2017](#); [Mohsen & Qassem, 2020](#); [Vandermeulen et al., 2024](#)), there is reason to assume that pausing behavior differs between learners who revise a text with or without feedback, but it is questionable whether the previous findings can be transferred to the context of text revision in combination with feedback.

For learners who revise without feedback, increased pausing could indicate higher cognitive engagement ([Révész et al., 2019](#)). For instance, a study by [de Smet et al. \(2018\)](#) showed that writing pauses related to revisions are longer compared to pauses when writing a new text, which suggests that cognitive processes during writing pauses differ depending on whether a learner is working on a first draft or a revision. However, receiving feedback for a revision could reduce cognitive load induced by a revision task which might be reflected in decreased pausing or in that the pauses are used differently (e.g., [Révész et al., 2019](#)). Accordingly, it can be assumed that

pausing behavior differs between learners who revise with and without feedback, however, we do not know enough about writing pauses in association with receiving instructional feedback to formulate clear expectations about their meaning in this context.

On the basis of the existing literature, we assume that increased pausing behavior is associated with higher cognitive engagement during text revisions among learners who revise without feedback (Lacruz et al., 2012; Medimorec & Risko, 2017; Mohsen & Qassem, 2020) and that the level of cognitive engagement—as well as behavioral engagement—is associated with learners' revision performance (Schiller et al., 2024). We assume that learners who receive feedback for a revision show different pausing behavior than those who revise without feedback, but we have no specific expectations regarding pausing patterns in relation to feedback.

It is important to keep in mind the limited interpretability of writing pauses. They could also indicate off-task behavior, similar to other time-related measures (Lindgren & Sullivan, 2006). Keeping this in mind, in the context of the present study, we refer to writing pauses as potential indicators of cognitive engagement to distinguish between cognitive and behavioral engagement (Wong et al., 2024). Analyzing writing pauses during the text revision of learners who receive automated feedback for the revision and of learners who revise without feedback could provide valuable insights into feedback processes, complementing previous research by attempting to operationalize cognitive learning engagement in the context of automated feedback and text revision.

3. The present study

The present study aimed to contribute to the existing literature by taking a process-oriented perspective on the effectiveness of automated feedback on learners' writing performance. Offline and online process measures reflecting learning engagement during text revision were used as indicators of behavioral learning engagement in relation to automated feedback. Additionally, different measures related to writing pauses were obtained as potential indicators of processes related to the cognitive dimension of learning engagement. On the one hand, the present study aimed to replicate previous findings obtained on the basis of offline measures used to assess behavioral engagement (i.e., edit distance and total time on task); on the other hand, it aimed to provide novel insights into the learners' role in the feedback process by using online measures obtained from keystroke logging to assess behavioral learning engagement (i.e., the total number of keystrokes, the typing time) and learners' pausing behavior (i.e., the total pausing time, the mean pause length, the number of pauses) during text revision. Studying feedback effects on text revision performance can provide information about the improvement of learners' performance on a specific task (Fleckenstein et al., 2023). We were also interested in the role that learning engagement plays in the ability of learners to apply feedback to new tasks; therefore, we implemented a subsequent writing task in addition to the text revision task (Fleckenstein et al., 2023). Considering learners' transfer performance can provide insights into whether the extent to which learners process automated feedback during a specific task (i.e., in the text revision) is also associated with their ability to transfer and apply the automated feedback to different tasks (Chan et al., 2024).

Aiming to replicate the results of Schiller et al. (2024), we addressed our first research question, using offline measures as indicators of learning engagement.

RQ1. *Do the edit distance and the total time on task, as indicators of behavioral engagement during text revision, mediate the expected effect of automated feedback on learners' revision performance and their performance on a transfer task?*

We implemented a between-subjects study design in which learners revised a writing task with or without feedback. With regard to the aim of replicating previous research and providing evidence on the robustness of the findings on learning engagement in this context, the edit distance and the total time on task (i.e., offline measures) were used to assess behavioral engagement during the text revision (Schiller et al., 2024). While the EFL learners in Schiller et al. (2024) were in Grades 7 to 9, the present study focused on more advanced EFL learners in Grade 10. Furthermore, the learners in the present study wrote an argumentative essay, whereas, in Schiller et al. (2024), the task was to write an email. The feedback in the present study was generated by a LLM. This means that the feedback in the present study was not reviewed by a teacher or another expert before it was provided to the learners. With regard to RQ1, we expected learners in the feedback condition to show higher behavioral engagement—measured using offline measures (i.e., edit distance and total time on task)—than learners in the control condition (RQ1.H1). We assumed that behavioral engagement, assessed with offline measures, would mediate the effect of automated feedback on the revision performance (RQ1.H1a; see Fig. 1). We also expected behavioral engagement assessed with offline measures to mediate the effect of automated feedback on learners' performance in a transfer task (RQ1.H1b).

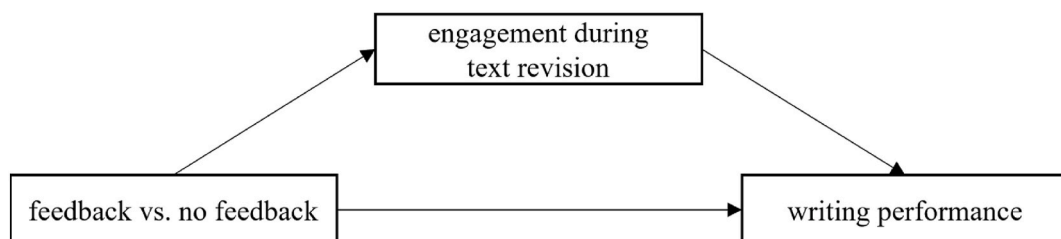


Fig. 1. Illustration of Expected Mediation

Note. The feedback effect on performance on a writing task (i.e., a revision and a transfer task) was expected to be mediated through learning engagement during text revision.

Aiming to extend previous research, we used online measures obtained from keystroke logging as indicators of learners' behavioral engagement and measures related to pausing behavior during the revision. Accordingly, our second research question was as follows.

RQ2. Do online measures obtained from keystroke logging (i.e., the total number of keystrokes and the typing time as indicators of behavioral learning engagement, and the total pausing time, the mean pause length, and the number of pauses) mediate the expected effect of automated feedback on learners' revision performance and their performance on a transfer task?

With regard to RQ2, we expected to find the same pattern of results as that expected regarding RQ1. We hypothesized that learners in the feedback condition would show higher behavioral learning engagement (RQ2.H2) during the text revision and that behavioral learning engagement would mediate the expected effect of the automated feedback on the learners' revision performance (RQ2.H2a). With regard to the writing pauses, we expected that the pausing behavior of learners in the feedback condition would differ from the pausing behavior of learners in the control condition (RQ2.H3). Following an explorative approach, we did not have specific assumptions about the exact patterns of pausing behavior in the conditions, but since writing pauses have been interpreted as potential indicators of different cognitive processes in writing (Lindgren & Sullivan, 2006), we assumed that the differences in pausing behavior would also mediate the expected effect of the automated feedback on the learners' revision performance (RQ2.H3a). Furthermore, we assumed that both behavioral learning engagement (RQ2.H2b) and the pausing behavior (RQ2.H3b) would mediate the effect of automated feedback on learners' transfer performance.

Our third research question (RQ3) was concerned with the variance in performance that is explained by the different indicators of behavioral engagement (i.e., offline and online measures) and was as follows.

RQ3. Do the indicators of behavioral engagement differ in how much variance they explain in learners' performance?

To answer RQ3, we exploratively examined the explanation of variance in revision performance scores by the different offline measures (i.e., edit distance and total time on task) and online measures (i.e., total number of keystrokes and typing time).

4. Methods

4.1. Acquisition and sample

The present article is based on secondary analyses of data from a larger research project (Meyer et al., 2024). The study was reviewed and approved by the Ministry of Education in the German federal state of Schleswig-Holstein and an independent ethics committee at the Leibniz Institute for Science and Mathematics Education. We contacted a selection of upper secondary schools in the German federal state of Schleswig-Holstein and invited them to participate in our study in the spring and summer of 2023. Schools contacted us on a voluntary basis and received further information and consent forms to be signed by the parents or guardians of the participating learners. The study was conducted during regular school hours and we collected data from 552 learners. The learners did not receive incentives to participate. We provided the schools with a report about the key findings of the study after the data collection was completed as a thank-you for their participation. We excluded 93 learners because of technical difficulties (Meyer et al., 2024). For the present study, data from six further learners were excluded from the analyses because they did not complete the first writing assignment, which served as a baseline measure. The final sample consisted of $N = 453$ EFL learners in the 10th grade ($M = 16.11$ years, $SD = 1.26$, 57 % female).

4.2. Design and procedure

We conducted a computer-based classroom study with a between-group design (feedback vs. no feedback), which we implemented in a digital survey tool (LimeSurvey 3). The participants were randomly assigned to the feedback condition ($n = 199$) or the control

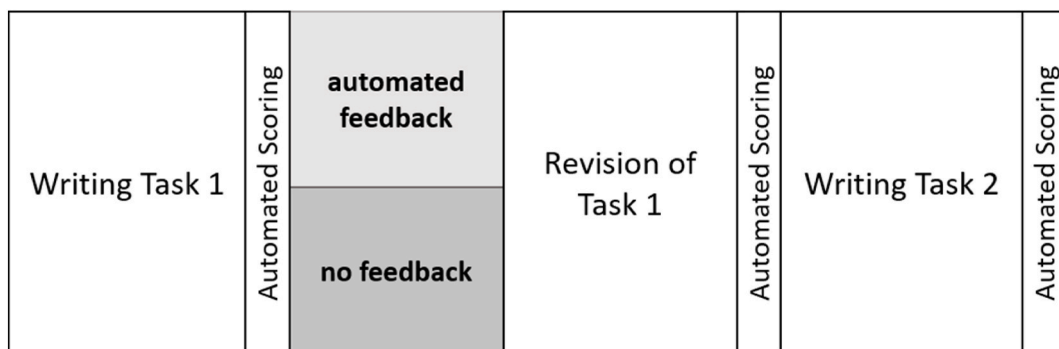


Fig. 2. Procedure

Note. We used offline measures (i.e., edit distance and total time on task) and online measures (i.e., total number of keystrokes, typing time, and writing pauses) that reflect learners' revision behavior.

condition ($n = 254$) within classes. All participants completed the first writing task. After submitting the first writing task, all participants were asked to revise their drafts either with automated feedback or without feedback, depending on the condition. After revising the first draft, all participants completed a subsequent writing task, which was based on a similar writing prompt and served as a transfer task. At the end of the test session, participants in both conditions were provided with a summative evaluation to compensate for the potential disadvantage of learners in the control condition. We used AWE rather than the LLM-generated feedback to provide the summative evaluation. The development of AWE-based feedback is preceded by several steps that involve didactic considerations (Rupp et al., 2019). This means that, at the time when our study was conducted, the accuracy of AWE-based summative evaluation was considered to be higher than that of an evaluation generated by an LLM (Steiss et al., 2024). The summative evaluation consisted of a holistic score ranging from 0 to 5, indicating the quality of the texts written by the learners (see Section 4.4.1, Writing Performance, for details of the scores). This summative evaluation thus ensured that learners in the control condition also received information about the quality of their texts, which they could use to evaluate their own texts and benefit from it. The maximum duration of a test session was 90 min. During this time, the learners completed the first writing task (20 min), the revision (no time limit), the transfer task (20 min), and several questionnaires (no time limit). Fig. 2 shows the procedure of the test session.

4.3. Materials

4.3.1. Technical equipment

We provided learners with tablets and keyboards for the test session and equipped the classroom with additional routers to ensure a stable internet connection. The learners accessed the online study in LimeSurvey 3 via a link on the tablet. The keystrokes during the text revision were recorded using a script that we implemented in the source code of the survey tool. For each keystroke event, the time, position, and key were recorded, including the backspace key. It was not possible to select characters, delete, insert or replace many characters with a single keystroke, or insert many characters at once. These restrictions were necessary in order to ensure that the position data were correct. Only keyboard activities that were entered into the text field in the survey tool were recorded.

4.3.2. Writing tasks

We adopted two writing tasks of comparable difficulty from the Test of English as a Foreign Language (TOEFL iBT®) for the first and second writing tasks (Keller et al., 2020). Learners were asked to write essays with regard to specific statements (i.e., “A teacher’s ability to relate well with learners is more important than excellent knowledge of the subject being taught” or “Television advertising directed toward young children aged two to five should not be allowed”). The order of the tasks was randomized. Learners were informed that a satisfactory essay should have at least 300 words and that they had a set time of 20 min to complete the task. The learners were not informed about the evaluation criteria, but the type of writing task should be familiar to students in the 10th grade in the German federal state of Schleswig-Holstein (Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein [Ministry of Education and Science Schleswig-Holstein], 2014). This means that learners can be expected to be familiar with the requirements of the task in terms of the evaluation criteria. During the 20 min, learners could not click the “Next” button. When the time was up, a message appeared and learners were automatically directed to the next page.

4.3.3. Revision

After completing the first writing task, learners were forwarded to the next page after approximately 1 min. We implemented the waiting time to avoid technical problems with the generation of automated feedback. To ensure that the waiting time did not confound the results, it was equal in both conditions. After the waiting time, the learners were asked to revise their drafts. They could edit their draft directly in the survey tool. During the revision, the writing task was displayed again in both conditions. Learners in the feedback condition were provided with automated feedback on content, structure, and language. Learners in the control condition did not receive feedback but were still instructed to revise their essay (“Please revise your text in the box below”). There was no time limit for the revision.

4.3.4. Automated scoring of text quality

For the automated scoring of text quality, we used a scoring model trained on a large text corpus and annotated by expert human raters who used a specific scoring rubric to assign a holistic score to the essays with regard to, for example, the organization of the essay, the provision of individual ideas, and the accurate use of the English language (Keller et al., 2020; Rupp et al., 2019). The algorithm was trained on the basis of linguistic features of the texts, including lexical, structural, and complexity-based attributes that are assumed to predict text quality (Zesch & Horbach, 2018). The algorithm was then used to score new texts and the training data were cross-validated. The model reached a quadratically weighted kappa of .76 on a held-out test set. A more detailed description of the scoring algorithm can be found in Meyer et al., 2024.

4.3.5. Automated feedback

The feedback provided to learners in the feedback condition was generated by an LLM (i.e., GPT-3.5 Turbo). We instructed the LLM to address the quality of the essay with regard to content, structure, and language, which are central aspects of argumentative essays (Keller et al., 2023). In developing the feedback prompt, we followed the principles for the design of effective feedback, ensuring that the feedback provided information on aspects that needed improvement and also provided examples of how these aspects could be addressed (Hattie & Timperley, 2007). We instructed the LLM to generate feedback appropriate for upper secondary EFL learners. We integrated a tabular format in the prompt and instructed the LLM to insert the information into the corresponding cells, with separate

columns containing hints and examples for improvement. The feedback on the three different aspects (i.e., content, structure, and language) was displayed simultaneously as one coherent feedback message above the text field in which the learners revised their drafts. The feedback was visible the whole time the text was being revised. The exact prompt and an example of the feedback can be found in Meyer et al. (2024).

4.4. Measures and operationalization

4.4.1. Writing performance

The learners' performance in the initial draft, the revision, and the transfer task was evaluated automatically by the algorithm described in Section 4.3.4, Automated Scoring of Text Quality, and in more detail in Meyer et al. (2024). Each of the three texts was assigned a holistic score between zero and five. A value of zero indicated low text quality, a value of five indicated high text quality.

4.4.2. Behavioral learning engagement

We used four different measures as indicators of behavioral engagement: the edit distance, the total time on task, the total number of keystrokes, and the typing time. The edit distance is a similarity measure based on the superficial similarity between two texts, here, the initial draft and the revision of the initial draft (Levenshtein, 1965). It indicates the number of changes made to the text, which become visible in the final version of the revision. The edit distance was calculated by using the Levenshtein algorithm. We normalized the edit distance to the text length to account for differences in text length so that it was comparable across texts of different lengths. The total time on task was retrieved from the log data, which were stored automatically by the survey tool. The total time on task is the duration from reaching the page with the revision task until clicking the forward button leading to the next page. As, in the feedback condition, both the feedback and the learner's draft were displayed for the entire time the learner spent on this page, the total time on task also includes the time spent reading the feedback.

The total number of keystrokes contains information about each keystroke made during the revision, including repeated substitutions or deletions of characters. We also obtained the typing time, which is the time from the first keystroke to the last keystroke made during the text revision, indicating how much time learners spent actively revising their essay.

4.4.3. Writing pauses

We retrieved three different measures related to learners' pausing behavior during the revision process: the total pausing time, the mean pause length, and the number of pauses. We calculated the mean inter keystroke interval (IKI) for each participant based on the keystrokes made during the initial writing task. The individual mean IKI (+1 standard deviation) was used as a minimum threshold to indicate a pause (de Smet et al., 2018). The mean individual pause threshold ($M = 0.88$ s, $SD = 0.61$) did not differ significantly between the feedback and the control condition (see Table S1 in the Supplementary Materials). In previous research using fixed pause thresholds, a minimum duration of 1 or 2 s has often been used to define pauses relevant for cognitive processes (e.g., Medimorec & Risko, 2017; Révész et al., 2017). However, for an experienced writer, a pause of 1 s might be long, while it is a short pause for a more inexperienced writer (e.g., Guo et al., 2018). Compared to using fixed pause thresholds, using individual pause thresholds takes into account individual differences in typing skills (de Smet et al., 2018; Roeser et al., 2024). We did not set an upper boundary to define the pauses, because we think that, in the context of text revision, longer pauses could be related to thinking about how to strengthen the argumentation or to plan how to improve the structure of the text. Therefore, we decided to include all pauses that exceeded the individual minimum threshold.

The total pausing time is the sum of all pauses made during the typing time (i.e., during the time between the first and the last keystroke). The mean length of the pauses was calculated based on all pauses that were made by a learner during the typing time, and the total number of pauses was the sum of pauses made during the typing time. We normalized the total pausing time and the number of pauses by the typing time to account for differences in total typing time. We used the typing time as a reference to normalize the variables because the total time on task contains the time learners spent reading the instructions and the feedback, which can be expected to be longer for learners in the feedback condition and would thus bias the pausing variables.

4.4.4. Covariates

We considered age, gender, participants' last English grade, and socioeconomic status (SES; i.e., parental education and number of books at home) as sociodemographic covariates. Grades in Germany range from 1 (very good) to 6 (failed). We inverted the English grade to obtain a variable where low values indicated low performance and high values indicated high performance. Parental education was assessed via a single-choice item with seven options, where we coded the lowest possible level of schooling (i.e., *did not attend school*) as one and the highest possible level of schooling (i.e., *school leaving certificate from an academic-track secondary school*) as seven (OECD, 2023). Furthermore, we controlled for initial performance by including the performance score on the first writing task as a covariate in the analyses; we also controlled for the order of the tasks.

4.5. Statistical analyses

4.5.1. Data preprocessing

We used R (Version 4.4.1) for statistical analyses (R Core Team, 2024). We inspected the distribution of the variables that served as indicators of behavioral engagement. The distributions of the edit distance and the total time on task were right-skewed. The distributions of the total number of keystrokes and the typing time were also right-skewed. The distribution of the normalized total

pausing time was left-skewed. The distributions of the mean pause length and the normalized number of pauses were right-skewed. We applied Box-Cox analyses and transformations using the optimal lambda values to achieve more normally distributed data and to stabilize the variance of the variables (Box & Cox, 1964; Osborne, 2010; Venables & Ripley, 2002).

4.5.2. Mediation analyses

We performed mediation analyses to test our hypotheses. The analyses were performed using the R package “lavaan” (Rosseel, 2012). Regarding the revision performance as an outcome variable, we performed seven separate mediation analyses, with the edit distance, the total time on task, the total number of keystrokes, the typing time, the total pausing time, the mean pause length, and the number of pauses serving as the mediators. We repeated the procedure with the transfer performance as the outcome variable.

For both outcomes, we first determined the total effect of feedback on performance (i.e., revision or transfer), controlling for the initial performance and the covariates. Then, we determined the direct effect of feedback on performance after including the mediator (i.e., edit distance, the total time on task, the total number of keystrokes, the typing time, the total pausing time, the mean pause length, or the number of pauses). We also determined the direct effect of feedback on behavioral learning engagement (i.e., edit distance, the total time on task and the total number of keystrokes) and the pause variables (total pausing time, the mean pause length and the number of pauses). We estimated the indirect effects and the portions of mediation based on 1000 bootstrapped samples (Tingley et al., 2014). In addition to the main analyses, we considered the four models with regard to the variance explained in the revision performance scores by the different indicators of behavioral engagement investigated.

5. Results

5.1. Descriptive statistics and bivariate correlations

Learners in the feedback condition and in the no-feedback control condition did not differ in their initial performance or socio-demographic covariates (see Table S1 in the Supplementary Materials). As can be seen in Table 1, there were no significant differences between conditions with regard to the text length of the initial draft, the revision, or the transfer task.

5.1.1. Group differences in revision behavior

Table 1 also displays group differences in revision behavior, revision performance and transfer performance. For the interpretation of the results, it should be noted that a total of 57 participants did not revise at all and that the portion of participants who made zero keystrokes did not differ significantly between conditions, $\chi^2 = 1.68$, $p = .195$. We found significant differences between learners in the feedback condition and the control condition with regard to the indicators of behavioral engagement during the text revision, with small to medium effect sizes. Learners who received feedback showed a significantly higher number of total keystrokes, a longer typing time, a higher edit distance, and a longer total time on task than learners who revised without feedback. With regard to the writing pauses, there was no significant difference between conditions with regard to the normalized total pausing time. The normalized number of writing pauses differed significantly between conditions, with a proportionally higher number of pauses among learners in the feedback condition. The mean pause length was significantly longer in the control condition compared to the feedback condition. This suggests that, in sum, learners who received automated feedback made proportionally more but shorter writing pauses during the revision than learners in the control condition.

Table 1
Group differences in revision behavior and revision and transfer performance.

Variable	Group		p (W) ^a	p (t) ^b	d	
	Complete Sample	Feedback ($n = 254$)				Control ($n = 199$)
	Mean (SD)	Mean (SD)				
1. Keystrokes	495.09 (466.07)	379.97 (410.26)	642.03 (492.05)	< .001	<.001	.51
2. Typing time	391.25 (288.31)	325.13 (276.05)	475.64 (282.17)	< .001	<.001	.41
3. Edit distance	319.44 (302.43)	392.03 (307.39)	262.56 (286.48)	< .001	.002	.30
4. Total time	525.08 (314.51)	625.64 (290.40)	446.29 (310.37)	< .001	<.001	.55
5. Pause length (seconds)	6.61 (6.52)	5.35 (5.01)	7.66 (7.39)	< .001	<.001	.37
6. Normalized total pausing time	.75 (.15)	.74 (.13)	.75 (.17)	.528	.496	.07
7. Normalized number of pauses	.66 (.12)	.68 (.10)	.65 (.12)	< .001	.002	.32
8. Revision performance	2.78 (0.90)	2.87 (0.92)	2.70 (0.88)	–	.016	.19
9. Transfer performance	1.52 (0.89)	1.59 (0.96)	1.46 (1.0)	–	.173	.13
10. Text length initial draft	1137.70 (455.50)	1137.85 (441.73)	1135.80 (466.87)	–	.962	.00
11. Text length revision	1408.48 (538.17)	1459.97 (576.68)	1367.91 (503.19)	–	.07	.17
12. Text length transfer task	982.47 (510.22)	1028.13 (480.28)	946.27 (530.93)	–	.09	.16

Note. Means of and differences in variables reflecting the revision behavior of learners in the feedback and in the control conditions. *SD* = standard deviation. We report Cohen's d as an estimated effect size.

^a Wilcoxon rank-sum tests were performed on the basis of nonnormally distributed variables before they were transformed.

^b Independent sample t -tests were performed after Box-Cox transformations were applied to variables and for normally distributed variables.

5.1.2. Correlations between engagement variables

Table S2 in the Supplementary Materials contains the bivariate correlations between all the main variables and covariates across the entire sample. As preliminary analyses, we also report the correlations separated by condition to facilitate the presentation of the results (see Table 2). These correlations should be considered when interpreting the findings of the present study. In both conditions, the four different indicators of behavioral engagement were significantly positively correlated with each other. This suggests that, among both learners who received feedback and learners who revised without feedback, those who spent more time on the revision also made more changes to their drafts. With regard to the pausing variables, we found the normalized total pausing time as well as the normalized number of pauses to be negatively correlated with the number of changes made to the text (i.e., keystroke or edit distance) in both conditions, suggesting that learners who paused more in total during the revision process revised their texts to a lower extent. The mean pause length of the individual pauses was negatively correlated with the changes made to the text (i.e., keystrokes) in both conditions, suggesting that making shorter writing pauses was associated with more changes made to the text. The mean pause length and the normalized number of pauses were negatively correlated with each other in the control condition, but they were not significantly correlated in the feedback condition. This suggests that, among learners who revised without feedback, more frequent pauses tended to be shorter in duration.

5.1.3. Correlations between engagement variables and performance per condition

With regard to the correlations between behavioral engagement and performance, we found different patterns for the feedback condition and the control condition. In the feedback condition, all of the four different indicators of behavioral engagement (i.e., edit distance, total time on task, total number of keystrokes, and typing time) were significantly positively correlated with both the revision performance and the transfer performance. In the control condition, we found significant positive correlations between the indicators of behavioral engagement, except for the edit distance, and the revision performance. We also found significant positive correlations between behavioral engagement and transfer performance in the control condition, but only with the number of keystrokes and the typing time as indicators of behavioral engagement. The pausing variables were not significantly correlated with the revision or the transfer performance.

5.2. Feedback effects on behavioral engagement and writing pauses (hypotheses H1, H2, & H3)

The feedback had a significant positive effect on behavioral learning engagement during the text revision with regard to both the offline measures (i.e., edit distance and total time on task) and the online measures (i.e., total number of keystrokes and typing time), supporting Hypotheses H1 and H2. This suggests that learners in the feedback condition made more changes to their draft during the revision and also spent more time on the revision than learners in the control condition. Table 3 contains details about the effects of feedback on behavioral engagement during the text revision.

We also found significant effects of the feedback on the writing pauses during the text revision, but there were inconsistencies in the direction of the effects, depending on the specific pause variable. Therefore, Hypothesis H3 was only partially supported. The feedback effect on the normalized total pausing time was not significant. The effect of the feedback on the mean pause length was negative, suggesting that receiving feedback for the text revision was associated with shorter writing pauses. The effect of feedback on the normalized number of pauses was positive, suggesting that learners who received feedback showed a proportionally higher number of writing pauses than learners in the control condition. Table S3 in the Supplementary Materials contains details about the effects of the feedback on the three variables related to typing pauses.

5.3. Mediation analyses

5.3.1. Mediation analyses with the revision performance as outcome (hypotheses H1a, H2a, & H3a)

We report the results of the full mediation models, which considered learners' initial performance and the sociodemographic covariates. Regarding behavioral engagement, we found a full mediation of the positive effect of the automated feedback on learners' revision performance with the offline measures (i.e., edit distance and total time on task), supporting Hypothesis H1a, and with the online measures (i.e., total number of keystrokes and typing time), supporting Hypothesis H2a. This means that, with regard to all of the four indicators, behavioral learning engagement during the text revision had a significant positive effect on performance on the revision, but the direct effect of the automated feedback on the revision was no longer significant when the indicator of behavioral learning engagement was included. The indirect effects were estimated to be significant, with significant estimated portions of mediation (POM) ranging from .42 to .88. Table 3 contains the detailed results, with revision performance as the outcome and behavioral engagement as the mediator.

Regarding the writing pauses (i.e., normalized total pausing time, mean pause length, and the normalized number of pauses), we did not find a mediation of the feedback effect on revision performance. Except for a small negative effect of the normalized total pausing time, the pause variables did not have a significant effect on revision performance, and the direct effect of feedback on revision performance did not change significantly when one of the pause variables was included. There were no significant indirect effects. Accordingly, Hypothesis H3a was not supported. Table S3 contains the detailed results of the mediation analyses with the revision performance as the outcome and cognitive engagement as the mediator.

5.3.2. Mediation analyses with the transfer performance as outcome (hypotheses H1b, H2b, & H3b)

We conducted mediation analyses with the transfer performance as the outcome, despite the missing total effect of automated

Table 2
Bivariate correlations per condition.

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Feedback Condition																
1. Keystrokes ^b	–															
2. Typing time ^b	.95***															
3. Edit distance ^{a,b}	.94***	.93***														
4. Total time ^b	.84***	.89***	.82***													
5. Normalized pausing time ^b	–.40***	–.15	–.42***	–.19*												
6. Pause length ^b	–.24**	–.04	–.13	–.05	.13											
7. Normalized number of pauses ^b	–.12	–.32***	–.25***	–.24***	.17*	–.57***										
8. Initial performance	.13	.09	–.09	.07	.06	–.02	.05									
9. Revision performance	.37***	.35***	.19**	.32***	–.07	–.10	–.03	.66***								
10. Transfer performance	.33***	.31***	.21**	.27***	.00	–.12	.06	.27***	.39***							
11. Gender	.22**	.23**	.18*	.20**	–.10	.03	–.11	.02	.18*	.28***						
12. Age	–.01	–.01	.00	–.04	–.07	.00	–.02	–.04	–.01	–.04	–.11					
13. English grade ^c	.14	.13	.04	.13	–.06	.03	.02	.32***	.25**	.32***	.08	.01				
14. Education father	.10	.11	.04	.13	.21**	–.14	.17*	.17*	.19**	.15*	–.04	–.12	.08			
15. Education mother	.01	.00	.01	.03	–.03	–.07	.08	.04	.07	.01	.06	–.06	.17*	.28***		
16. Books	.25***	.25***	.23**	.24***	.04	.04	.01	.08	.19*	.12	.01	–.08	.04	.26***	.26***	
17. Task order	.01	–.03	–.04	.00	.05	–.11	.05	.25***	.23**	–.03	.00	–.15*	–.08	.14*	.12	–.02
Control Condition																
1. Keystrokes ^b	–															
2. Typing time ^b	.92***															
3. Edit distance ^{a,b}	.87***	.77***														
4. Total time ^b	.77***	.84***	.70***													
5. Normalized pausing time ^b	–.33***	.09	–.37***	.02												
6. Pause length ^b	–.37***	–.04	–.33***	–.04	.39***											
7. Normalized number of pauses ^b	.13	–.15*	.14*	–.08	–.16*	–.53***										
8. Initial performance	.08	.14*	–.20**	.06	.23***	–.02	–.06									
9. Revision performance	.32***	.36***	.08	.34***	.11	–.08	.02	.64***								
10. Transfer performance	.14*	.14*	.02	.12	.11	–.08	–.04	.27***	.31***							
11. Gender	.12	.12	.04	.07	.01	–.05	–.04	.21**	.12	.16*						
12. Age	–.07	–.08	.03	–.09	–.08	–.08	.15*	–.18**	–.21***	–.09	.00					
13. English grade ^c	.24***	.19**	.13*	.18**	–.14*	–.17*	.15*	.19**	.28***	.35***	.20**	.03				
14. Education father	.10	.14*	–.09	.12	.02	.01	–.10	.26***	.29***	.11	–.07	–.45***	.16*			
15. Education mother	.04	.09	–.12	.11	.09	.00	–.11	.24***	.28***	.17*	.00	–.32***	.16*	.71***		
16. Books	.09	.09	.00	.05	.12	–.10	.01	.17**	.21***	.17*	.07	–.04	.21***	.29***	.30***	
17. Task order	–.02	.04	.04	.05	.06	.04	–.03	.09	.14*	–.17**	.06	.08	–.05	–.06	–.08	.02

* $p < .05$. ** $p < .01$. *** $p < .001$.

^a Normalized by text length.

^b Box-Cox transformations were applied.

^c Inverted variable (here, 1 = failed, 6 = very good)

Table 3

Mediation analyses with the revision performance as outcome and behavioral engagement as the mediator.

Variable	Engagement Indicator											
	Edit Distance ^a			Total Time on Task			Number of Keystrokes			Typing Time		
	β	SE	95 % CI	β	SE	95 % CI	β	SE	95 % CI	β	SE	95 % CI
Total Effect (identical for all four analyses)												
Outcome: Revision Performance												
Feedback vs. control	.10*	.04	[.02, .17]	.10*	.04	[.02, .17]	.10*	.04	[.02, .17]	.10*	.04	[.02, .17]
1st writing score	.60***	.05	[.51, .70]	.60***	.05	[.51, .70]	.60***	.05	[.51, .70]	.60***	.05	[.51, .70]
Gender	.05	.04	[-.03, .13]	.05	.04	[-.03, .13]	.05	.04	[-.03, .13]	.05	.04	[-.03, .13]
Age	-.05	.04	[-.13, .03]	-.05	.04	[-.13, .03]	-.05	.04	[-.13, .03]	-.05	.04	[-.13, .03]
English grade ^b	.11*	.04	[.03, .19]	.11*	.04	[.03, .19]	.11*	.04	[.03, .19]	.11*	.04	[.03, .19]
Education father (SES)	.06	.05	[-.04, .15]	.06	.05	[-.04, .15]	.06	.05	[-.04, .15]	.06	.05	[-.04, .15]
Education mother (SES)	.02	.05	[-.07, .12]	.02	.05	[-.07, .12]	.02	.05	[-.07, .12]	.02	.05	[-.07, .12]
Number of books at home (SES)	.09*	.04	[.02, .17]	.09*	.04	[.02, .17]	.09*	.04	[.02, .17]	.09*	.04	[.02, .17]
Task order	.09*	.04	[.02, .17]	.09*	.04	[.02, .17]	.09*	.04	[.02, .17]	.09*	.04	[.02, .17]
Direct Effects												
Outcome: Revision Performance												
Feedback vs. control	.05	.04	[-.02, .13]	.01	.04	[-.06, .09]	.02	.04	[-.06, .09]	.03	.04	[-.04, .11]
1st writing score	.65***	.05	[.56, .74]	.61***	.04	[.52, .70]	.60***	.04	[.52, .69]	.60***	.04	[.51, .68]
Gender	.03	.04	[-.05, .11]	.03	.04	[-.05, .10]	.02	.04	[-.05, .09]	.02	.04	[-.05, .09]
Age	-.04	.04	[-.12, .05]	-.03	.03	[-.09, .03]	-.03	.04	[-.10, .04]	-.03	.03	[-.10, .04]
English grade ^b	.09*	.04	[.01, .16]	.08	.04	[.00, .15]	.07*	.04	[.00, .15]	.09*	.04	[.01, .16]
Education father (SES)	.06	.05	[-.03, .15]	.03	.04	[-.05, .12]	.05	.04	[-.04, .14]	.04	.04	[-.04, .13]
Education mother (SES)	.05	.05	[-.05, .14]	.03	.04	[-.05, .12]	.04	.05	[-.05, .13]	.04	.04	[-.05, .12]
Number of books at home (SES)	.06	.04	[-.02, .14]	.07	.04	[-.01, .15]	.06	.04	[-.02, .13]	.05	.04	[-.02, .13]
Task order	.08*	.04	[.00, .15]	.08	.04	[.01, .15]	.08*	.04	[.01, .15]	.08*	.04	[.01, .15]
Engagement	.24***	.04	[.17, .31]	.28***	.03	[.22, .35]	.29***	.04	[.22, .36]	.29***	.03	[.22, .35]
Outcome: Engagement												
Feedback vs. control	.16***	.05	[.07, .25]	.28***	-.05	[.19, .37]	.25***	.05	[.16, .34]	.20***	.05	[.11, .29]
1st writing score	-.21***	.06	[-.32, -.09]	-.02	.05	[-.12, .08]	-.01	.06	[-.12, -.10]	.02	.06	[-.09, .13]
Gender	.10*	.05	[.00, .20]	.11	.05	[.01, .20]	.13*	.05	[.03, .23]	.14**	.05	[.04, .24]
Age	-.02	.06	[-.14, .09]	-.05	.06	[-.16, .07]	-.04	.06	[-.15, .07]	-.04	.06	[-.15, .07]
English grade ^b	.12*	.05	[.01, .22]	.13*	.05	[.03, .23]	.16**	.05	[.05, .26]	.12*	.05	[.02, .23]
Education father (SES)	.00	.07	[-.13, .13]	.09	.07	[-.05, .22]	.04	.06	[-.08, .16]	.06	.06	[-.06, .19]
Education mother (SES)	-.10	.07	[-.23, .03]	-.03	.06	[-.14, .09]	-.07	.06	[-.19, .05]	-.06	.06	[-.18, .06]
Number of books at home (SES)	.14*	.05	[.03, .24]	.09	.05	[-.01, .19]	.13*	.05	[.03, .23]	.13*	.06	[.03, .24]
Task order	.04	.05	[-.05, .13]	.04	.04	[-.06, .13]	.01	.05	[-.08, .10]	.02	.54	[-.08, .11]
Estimation of the Indirect Effect (Full Model)												
	β	SE	95 % CI	β	SE	95 % CI	β	SE	95 % CI	β	SE	95 % CI
Indirect Effect	.04**	.01	[.01, .06]	.08***	.02	[.05, .11]	.07***	.02	[.04, .10]	.06***	.02	[.03, .09]
POM	.42*	.20	[.02, .82]	.88*	.38	[.14, 1.62]	.83*	.38	[.10, 1.57]	.63*	.28	[.08, 1.9]

* $p < .05$. ** $p < .01$. *** $p < .001$.^a Normalized against text length.^b Inverted variable (here, 1 = failed, 6 = very good).

feedback on transfer performance (Hayes, 2022; Zhao et al., 2010). We report the direct effect of the indicators of behavioral engagement and the pause variables on the transfer performance from the full model, considering the initial performance and the covariates. We found significant direct effects of behavioral engagement on learners' transfer performance both with the offline measures (i.e., edit distance and total time on task) and the online measures of behavioral engagement during the text revision (i.e., total number of keystrokes and typing time). The direct effect of the automated feedback on the transfer performance was not significant and the effect size decreased slightly when the indicators of behavioral learning engagement during the text revision were included. We found significant indirect effects with regard to the four mediators, supporting Hypotheses H1b and H2b. The estimated POMs regarding the four indirect effects were not significant and should not be interpreted because of the small direct effect of the automated feedback (Preacher & Kelley, 2011). Table 4 contains the detailed results of the mediation analyses with the transfer performance as the outcome and behavioral engagement as the mediator.

Regarding the writing pauses, we found neither significant effects of the normalized total pausing time, the mean pause length, or the normalized number of pauses on the transfer performance nor significant indirect effects related to the feedback. Therefore, there was no support for Hypothesis H3b. Table S4 contains the detailed results of the mediation analyses with the transfer performance as the outcome and the variables related to typing pauses as the mediators.

5.4. Explanation of variance in revision performance

We examined the amount of variance in the revision performance scores that was explained by the different models with the offline

Table 4

Mediation analyses with the transfer performance as outcome and behavioral engagement as the mediator.

Variable	Engagement Indicator											
	Edit Distance ^a			Total Time on Task			Total Time on Task			Typing Time		
	β	SE	95 % CI	β	SE	95 % CI	β	SE	95 % CI	β	SE	95 % CI
Total Effect (identical for all four analyses)												
Outcome: Transfer Performance												
Feedback vs. control	.06	.05	[-.03, .15]	.06	.05	[-.03, .15]	.06	.05	[-.03, .15]	.06	.05	[-.03, .15]
1st writing score	.21***	.05	[.11, .31]	.21***	.05	[.11, .31]	.21***	.05	[.11, .31]	.21***	.05	[.11, .31]
Gender	.16**	.05	[.06, .26]	.16**	.05	[.06, .26]	.16**	.05	[.06, .26]	.16**	.05	[.06, .26]
Age	-.03	.04	[-.11, .04]	-.03	.04	[-.11, .04]	-.03	.04	[-.11, .04]	-.03	.04	[-.11, .04]
English grade ^b	.25***	.05	[.15, .36]	.25***	.05	[.15, .36]	.25***	.05	[.15, .36]	.25***	.05	[.15, .36]
Education father (SES)	.03	.07	[-.10, .15]	.03	.07	[-.10, .15]	.03	.07	[-.10, .15]	.03	.07	[-.10, .15]
Education mother (SES)	-.01	.06	[-.13, .11]	-.01	.06	[-.13, .11]	-.01	.06	[-.13, .11]	-.01	.06	[-.13, .11]
Number of books at home (SES)	.08	.05	[-.02, .18]	.08	.05	[-.02, .18]	.08	.05	[-.02, .18]	.08	.05	[-.02, .18]
Task order	-.14*	.04	[-.23, -.05]	-.14*	.04	[-.23, -.05]	-.14*	.04	[-.23, -.05]	-.14*	.04	[-.23, -.05]
Direct Effects												
Outcome: Transfer Performance												
Feedback vs. control	.04	.05	[-.05, .13]	.02	.05	[-.06, .13]	.02	.05	[-.08, .11]	.02	.05	[-.06, .11]
1st writing score	.23***	.05	[.13, .33]	.21***	.05	[.11, .31]	.21***	.05	[.11, .31]	.20***	.05	[.11, .30]
Gender	.15**	.05	[.05, .25]	.15**	.05	[.05, .24]	.14**	.05	[.04, .24]	.14**	.05	[.04, .24]
Age	-.03	.04	[-.11, .05]	-.03	.04	[-.10, .05]	-.02	.04	[-.10, .05]	-.02	.04	[-.10, .05]
English grade ^b	.24***	.05	[.14, .35]	.24***	.05	[.14, .34]	.23***	.05	[.13, .34]	.24***	.05	[.14, .34]
Education father (SES)	.01	.06	[-.10, .15]	.02	.06	[-.10, .14]	.02	.06	[-.10, .15]	.02	.06	[-.10, .14]
Education mother (SES)	.00	.06	[-.12, .12]	-.01	.06	[-.12, .11]	.00	.06	[-.11, .12]	.00	.06	[-.12, .11]
Number of books at home (SES)	.06	.05	[-.04, .16]	.07	.05	[-.03, .17]	.06	.05	[-.04, .16]	.06	.05	[-.04, .16]
Task order	-.15**	.04	[-.24, -.06]	-.15	.04	[-.23, -.06]	-.15**	.04	[-.23, -.06]	-.15**	.04	[-.23, -.06]
Engagement	.13**	.05	[.04, .22]	.12*	.06	[.01, .23]	.16**	.05	[.06, .25]	.16**	.05	[.06, .25]
Outcome: Engagement												
Feedback vs. control	.16***	.05	[.07, .25]	.28***	.05	[.18, .37]	.25***	.05	[.16, .34]	.20***	.05	[.11, .29]
1st writing score	-.20**	.06	[-.32, -.09]	-.02	.05	[-.12, .08]	-.01	.06	[-.12, .10]	.02	.06	[-.09, .13]
Gender	.11*	.05	[.01, .21]	-.11*	.05	[.02, .20]	.13*	.05	[.04, .29]	.14**	.05	[.04, .24]
Age	-.02	.06	[-.14, .09]	-.05	.06	[-.16, .07]	-.04	.06	[-.15, .07]	-.04	.06	[-.15, .07]
English grade ^b	.11*	.05	[.01, .22]	.12*	.05	[.03, .22]	.15**	.05	[.05, .25]	.12*	.05	[.01, .22]
Education father (SES)	.00	.07	[-.13, .13]	.08	.07	[-.05, .22]	.04	.06	[-.08, .16]	.06	.06	[-.07, .19]
Education mother (SES)	-.10	.07	[-.22, .03]	-.03	.06	[-.14, .09]	-.07	.06	[-.19, .05]	-.06	.06	[-.18, .06]
Number of books at home (SES)	.14*	.05	[.03, .24]	.09	.05	[-.01, .19]	.13*	.05	[.013, .23]	.13*	.06	[.02, .24]
Task order	.04	.05	[-.05, .13]	.04	.05	[-.06, .13]	.01	.05	[-.08, .10]	.01	.05	[-.08, .11]
Estimation of the Indirect Effect (Full Model)												
	β	SE	95 % CI	β	SE	95 % CI	β	SE	95 % CI	β	SE	95 % CI
Indirect Effect	.02*	.01	[.00, .04]	.03*	.02	[.00, .07]	.04*	.01	[.01, .07]	.03*	.01	[.01, .05]
POM	.34	.32	[-.27, .96]	.60	.59	[-.55, 1.74]	.72	.64	[-.54, 1.98]	.57	.50	[-.42, 1.55]

* $p < .05$. ** $p < .01$. *** $p < .001$.^a Normalized against text length.^b Inverted variable (here, 1 (equal sign) failed, 6 (equal sign) very good).

and the online measures of behavioral engagement, including edit distance ($R^2 = .46$), the total number of keystrokes ($R^2 = .46$), the total time on task ($R^2 = .48$), and the typing time ($R^2 = .49$) as predictors. Thus, the amount of variance explained by the models that included the edit distance and the total number of keystrokes was the same. The model that included the typing time explained 1 % more variance in the revision scores than the model including the total time on task. The total time on task and the active typing time explained 2 %–3 % more variance in the revision scores than the edit distance or the total number of keystrokes.

6. Discussion

The aim of the present study was to investigate automated feedback effectiveness from a process-oriented perspective. We replicated previous research using offline measures as indicators of behavioral engagement and we used additional online measures obtained from keystroke logging that capture text revision behavior in an even more differentiated way. The present study offers novel insights into learning engagement as a decisive underlying process for the effectiveness of automated feedback. By using online measures to operationalize behavioral learning engagement and to analyze writing pauses as potential indicators of cognitive engagement, this study contributes to the existing body of research on automated feedback effectiveness. In the following, we will first discuss the results regarding behavioral learning engagement and then proceed with the results regarding the writing pauses.

With regard to our first research question (RQ1), we found that, on the basis of the offline measures of behavioral engagement (i.e., edit distance and total time on task), learners who received automated feedback showed higher levels of behavioral learning engagement during the text revision. That is, aligning with previous findings (Schiller et al., 2024), our results showed that they made

more changes to their text during the revision (i.e., had a higher edit distance) and spent more time on the revision (i.e., had a longer total time on task) than learners who did not receive feedback. With regard to the text revision performance, we were able to replicate previous findings that suggested that behavioral learning engagement mediates substantial portions of the effects of automated feedback on learners' text revision performance (Schiller et al., 2024), supporting Hypothesis H1a.

Regarding our second research question (RQ2), we found the same patterns of results for the online measures of behavioral engagement (i.e., total number of keystrokes and typing time), providing further evidence on the role of behavioral learning engagement in feedback effectiveness, supporting Hypotheses H2 and H2a. Our findings regarding the relationship between writing behavior and revision performance align with a study by Zhang et al. (2019) presented above. Zhang et al. (2019) found a longer typing time and a higher number of keystrokes to be associated with higher essay scores, emphasizing the importance of learners' behavioral engagement during the writing process (Zhang et al., 2019). The results of the present study illustrate how automated feedback can support learners in engaging in writing processes that are associated with improved writing performance (Zhang et al., 2019).

Regarding the transfer task, we found that higher levels of behavioral engagement during the revision of the first writing task were associated with essays of higher quality in the transfer task. The total effect of the automated feedback on learners' performance on the transfer task was not significant, but we found significant indirect effects with regard to all of the four indicators of behavioral engagement (i.e., both offline and online measures). The POMs of the mediation effects with regard to the transfer performance were not significant. Interpreting the POM as an effect size of a mediation can be misleading if the total effect is close to zero and not significant (Preacher & Kelley, 2011). Our findings suggest that the association between automated feedback and learners' performance on the transfer task was mediated by their behavioral engagement during the revision of the initial draft, supporting Hypotheses H1b and H2b. This aligns with findings by Fleckenstein et al., 2024, who found a significant indirect effect of the time learners spent reading automated feedback (i.e., behavioral engagement with feedback) on their writing performance. The findings of Fleckenstein et al., 2024 and the present study suggest that learners only benefit from automated feedback in the long term—in terms of performance on subsequent writing tasks—if they show high levels of engagement with the feedback they receive (Fleckenstein et al., 2024), as well as during the revision of the text on which they received automated feedback. This highlights the importance of placing more emphasis on the proactive role of the learner in the effectiveness of automated feedback (Winstone & Nash, 2023). Further research using longitudinal study designs is needed to investigate the role of learning engagement in the transfer of feedback effects.

The comparisons of learners' pausing behavior between the feedback condition and the control condition (RQ2) revealed some interesting insights. The inconsistent effects of the feedback on the different pausing variables suggest that learners who received feedback paused proportionally more than learners in the control condition, but the pauses were shorter; this means that Hypothesis H3 was only partially supported. In line with this, we found a small but significant negative effect of the feedback on the pause length and a small but significant positive effect of the feedback on the number of pauses. However, the estimated indirect effects related to the writing pauses were not significant and we did not find a mediation effect on revision or transfer performance; thus, Hypotheses H3a and H3b were not supported either.

Previous studies from the field of writing research have suggested that more frequent and longer pauses indicate higher cognitive effort, which is related to cognitive engagement (e.g., Lacruz et al., 2012; Medimorec & Risko, 2017; Mohsen & Qassem, 2020). However, most of the existing studies did not investigate writing pauses in the context of text revision with feedback and it is possible that, depending on the context and the type of writing task, less pausing or shorter pauses do not necessarily mean that learners are not cognitively engaged. Writing pauses might have completely different meanings when investigated in association with feedback. On the one hand, it is possible that learners in the feedback condition paused more frequently because they switched between their draft and the feedback and then proceeded writing after a short typing pause to incorporate what they read in the feedback message (Bolzer et al., 2015). In other words, the feedback might have been associated with reduced cognitive load because it delivered starting points and ideas for the revision, which then resulted in shorter individual pauses during the revision process (Révész et al., 2017). On the other hand, the individual pauses that were longer among learners in the control condition could indicate phases of intensive thinking and reflecting (Lindgren & Sullivan, 2006; Vandermeulen et al., 2023). However, in general, it is also possible that the pauses were related to off-task behavior, which would explain why they were not associated with performance (Lindgren & Sullivan, 2006).

In sum, the present findings confirm that cognitive engagement is more complex and difficult to measure than behavioral engagement (Wong et al., 2024). It becomes clear that, as already pointed out by Wong et al. (2024), operationalizing the cognitive dimension of learning engagement is particularly challenging, even in the context of writing, where keystroke logging data provides detailed information about processes at the task level. It is important to emphasize that using the pausing variables was an attempt to assess processes that take place during text revision and might reflect cognitive learning engagement in dependence of feedback at the task level. When interpreting the results, it has to be kept in mind that cognitive processes related to writing and revising do not necessarily take place only during typing pauses. It is also possible that learners think and plan in parallel to showing proactive typing behavior (Roeser et al., 2025), or that cognitive processes are reflected in certain manifest keystroke patterns rather than only in inactive phases of the writing process (e.g., Hall et al., 2024; Xu, 2024). In the context of the present study, it is not possible to provide a clear interpretation regarding the writing pauses as potential indicators of cognitive learning engagement. However, if they do not reflect cognitive engagement, they still carry some information about internal processes. Therefore, further research is needed to investigate how exactly cognitive engagement manifests in the context of text revision and in relation to receiving feedback.

With regard to our third research question (RQ3), we did not find evidence of online measures of behavioral engagement explaining more variance in writing performance than offline measures. However, we found that both measures related to the time learners spent revising (i.e., the total time on task and the typing time) explained more variance in learners' revision performance scores than the extent to which learners revised their drafts did (i.e., the edit distance or the total number of keystrokes). This suggests that, in the context of writing, time on task measures might contain information about cognitive processes such as planning or reflecting that do

not become visible in learners' observable text revision behavior. The use of online measures to investigate learning engagement on a fine-grained level should be further explored in future research.

6.1. Implications

Aligning with previous research (e.g., Schiller et al., 2024; Zhang & Deane, 2015; Zhang et al., 2019), the present study shows that learning engagement is crucial for the development and improvement of learners' writing performance. It shows that receiving (automated) feedback on writing assignments is effective because it can increase learners' behavioral engagement during text revisions. We investigated this relationship in the context of automated feedback. Given that we considered the principles for designing effective feedback (Hattie & Timperley, 2007), we would expect the findings to hold in the context of feedback from other sources. Feedback does not simply deliver information about the current and the target state of a task a learner is working on. It effectively supports learners in engaging in a task to a higher extent. Especially in light of the increasing use of digital tools, such as automated and AI-generated feedback, the inclusion of learners' behavior in the investigation of learning effects is becoming increasingly relevant and should be considered in future research (Sailer et al., 2024). For instance, the question of whether certain types of feedback are related to higher learning engagement compared to other types of feedback should be investigated (Fleckenstein et al., 2024). To do this, future research should follow interdisciplinary approaches by combining automated feedback systems and process measures. The present study provides an example of how such an approach can be used to better explain learners' responses to feedback instead of only considering performance outcomes. Further studies are needed to gain a more precise picture of revision processes and to clarify how online measures can contribute to a more complete understanding of learning engagement in relation to automated feedback.

6.2. Limitations and future directions

The present study has some limitations that need to be considered. First, the accuracy of the measurement of the text quality (i.e., automatically assigned performance scores) is limited. Errors in the automated scoring of the text quality cannot be completely ruled out, which means that the performance scores for the three texts (i.e., initial draft, revision, and transfer task) may have contained measurement errors. Furthermore, the scoring algorithm assigned a holistic score to each text, and no further differentiation can be made with regard to more detailed characteristics of the texts, for instance, the complexity. The limited accuracy of the text quality scoring might have led to an underestimation of the effects, which may have contributed to the small effects of feedback and engagement on performance that we found in the present study.

Second, there are limitations in relation to the automated feedback. The present study was conducted at a time when the quality of LLM-generated feedback had not been studied much. Since then, studies comparing the quality of LLM-generated feedback with expert feedback found that the quality and perceived usefulness of LLM-generated feedback on written texts was inferior to expert feedback (Jansen et al., 2024; Steiss et al., 2024), although it has been emphasized that LLMs can provide relatively satisfactory feedback, given the immediacy of the feedback provision (Steiss et al., 2024). In the present study, we did not examine the quality of the feedback provided to learners. Therefore, we cannot exclude the possibility that the LLM hallucinated and included incorrect information in the feedback messages (Pardos & Bhandari, 2024). Furthermore, the design of the present study did not allow for a more differentiated analysis of engagement in connection with individual aspects of the automated feedback, as learners were presented with one coherent feedback message. Future studies should include qualitative analyses of automated feedback to provide information about the reliability of feedback, the fit between the feedback and learners' texts, and potential differences in the relationship between feedback, learning engagement, and performance with regard to specific aspects of feedback.

Third, we did not find that writing pauses were related to performance or that they played a significant role in mediating feedback effectiveness. The absence of these effects may be due to limitations in our measures and our conceptualization of writing pauses. We defined pauses as inactive time during the typing process (i.e., pauses between keystrokes) exceeding an individual minimum pause threshold (de Smet et al., 2018; Roeser et al., 2025). Referring to pause thresholds that are based on the individual learners' writing behavior instead of using fixed thresholds allows for the consideration of differences in the writing experience of learners (de Smet et al., 2018). However, in the present study, we did not implement a copy task to examine learners' "baseline" writing skills (e.g., Schuurman et al., 2022), so we used the keystroke logging data of the initial writing task to calculate the individual pause thresholds. This is a methodological limitation, because learners' writing behavior during tasks that require the generation of text could be influenced not only by learners' typing experience, but also by prior knowledge about the topic of the writing prompt or their experience with writing texts of a specific genre (Roeser et al., 2024). Future studies should make sure to include a separate copy task to be able to examine learners' typing skills as neutral as possible. Next to the methodological limitations, the present study also comes with limitations regarding the interpretability of writing pauses. Analyzing writing pauses was an attempt to assess potential indicators of processes related to the cognitive dimension of learning engagement. However, it is important to emphasize that the meaning of pauses probably varies depending on the context, such as text revision with or without (de Smet et al., 2018; Révész et al., 2019).

Fourth, in connection with the previous point, the exact way in which writing pauses are used by learners and the type of cognitive processes that occur during the pauses may vary across different populations. The present study is based on a sample of EFL learners in the 10th grade. It is possible that this group of writers still encounters difficulties with writing in English and, therefore, uses pauses for different cognitive processes than more experienced writers would (e.g., de Smet et al., 2018; Garcés-Manzanera, 2023; Lindgren & Sullivan, 2006). Furthermore, learners who are more experienced with typing might potentially show more consistent typing speed and take fewer long pauses while writing (Guo et al., 2018). Future studies should take such potential differences into account, particularly when it comes to supporting foreign-language writing skills through automated feedback tools.

Fifth, similar to the writing pauses (Lindgren & Sullivan, 2006), the two indicators related to the revision time (i.e., total time on task and typing time) as well as the writing pauses are difficult to interpret. We believe that the typing time that we obtained from the keystroke logging data contains information relevant for behavioral engagement that remains hidden when only the total time on task is considered. However, due to our operationalization of the typing time (i.e., the duration from the first to the last keystroke), it is possible that one learner, who revised an entire text, had the same typing time as another learner, who changed only two characters with a long break in between the two characters. Despite the additional information that we received from the edit distance and the total number of keystrokes, as well as from the writing pauses, we cannot make inferences about how the learners used the total time or the typing time when they did not proactively edit their draft. We cannot rule out the possibility that the time-related measures (i.e., total time on task, typing time and writing pauses) are associated with off-task behavior (Kovanovic et al., 2016; Lindgren & Sullivan, 2006). In the context of measuring psychological mechanisms related to learning (here, behavioral engagement related to automated feedback), using process measures is a novel approach that has not yet been validated with multimodal data sources (e.g., Goldhammer et al., 2021). To be able to interpret certain process measures, like writing pauses, future studies should use multimodal approaches, for instance, by using eye tracking or think-aloud protocols in combination with keystroke logging to retrieve information about what learners actually do during the pauses (Bolzer et al., 2015; Bouwer & Dirks, 2023; Fan et al., 2023; Révész et al., 2017, 2019), or by comparing behavioral indicators of engagement during a learning activity to data collected with self-reports explicitly developed to assess task-related engagement (Choi et al., 2023; Greene & Miller, 1996; Quick et al., 2020; Smiley & Anderson, 2011).

Sixths, building on the previous points, it is important to emphasize that the behavioral and the cognitive dimension of learning engagement are often difficult to distinguish (Wong et al., 2024). The boundaries between proactive observable behavior and cognitive processes are blurred, especially in writing, where thinking and planning can occur in parallel to observable typing (Roeser et al., 2025). Certain patterns in writing behavior—including both observable keystrokes and writing pauses—might be associated with specific phases of the writing process, and it is possible that observable keystrokes do not only indicate behavioral engagement, but also contain information about cognitive processes (Hall et al., 2024; Xu, 2024).

Moreover, sevenths, the quantitative operationalization of learning engagement in the current study does probably not fully capture the complexity of learning engagement. Future feedback research would profit from combining quantitative and qualitative approaches to investigate learning engagement in the context of automated feedback in a more nuanced way, for instance, by analyzing writing patterns in more detail. For example, previous studies by Zhang et al. (2019) and by Talebinamvar and Zarrabi (2022) analyzed learners' writing processes and found similar associations between a certain writing pattern and a certain writing performance. In their studies, learners who paused at the beginning and then showed a fast and consistent text production tended to perform better than learners with different writing patterns (Talebinamvar & Zarrabi, 2022; Zhang et al., 2019). Considering that many learners do not benefit from automated feedback because they do not engage with feedback at all (Bahr et al., 2025; Meyer et al., 2025), more in-depth insights into how individual learners interact with automated feedback during text revision processes would be valuable for the development and improvement of feedback tools that aim to increase individual learners' engagement and support them in their writing in a more needs-oriented way.

7. Conclusion

Following recent calls to take a process-oriented perspective in order to gain a more complete understanding of the effectiveness of (automated) feedback, we focused on learners' engagement during text revision and its role in the effect of LLM-generated feedback in the context of writing. Focusing on the use of different process measures (i.e., offline and online measures) as indicators of learning engagement during text revision, the present study contributes to the existing literature by providing further evidence on the role that learning engagement plays in the effectiveness of automated feedback. Automated feedback helps learners engage with a task, such as revising a text, and their behavioral engagement is—in turn—associated with better performance. We used writing pauses as potential indicators of cognitive engagement to provide a more comprehensive picture of learning engagement. We found differences in cognitive engagement between learners in the feedback condition and the control condition, however, cognitive engagement, as we conceptualized it, did not seem to be decisive for the effectiveness of automated feedback on performance. Further research is needed to examine and validate online measures created to capture the different dimensions of learning engagement (Goldhammer et al., 2021; Henrie et al., 2015).

The present study highlights potential pathways for future research, utilizing keystroke logging to obtain detailed, task-specific measures of learning engagement. Further studies using different types of writing tasks, different criteria of text quality, and different study populations are needed to provide an empirical basis for drawing robust conclusions about the role of learning engagement in the effectiveness of automated feedback on writing.

CRedit authorship contribution statement

Ronja Schiller: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Johanna Fleckenstein:** Writing – review & editing, Funding acquisition, Conceptualization. **Lars Höft:** Writing – review & editing, Methodology. **Andrea Horbach:** Writing – review & editing, Methodology. **Jennifer Meyer:** Writing – review & editing, Conceptualization.

Author note

This work was funded by the German Federal Ministry of Education and Research grant number 01JG2104.

Registration

This study was not preregistered.

Permission to reproduce material from other sources

Not applicable.

Declaration of competing interest

We have no conflict of interest to disclose.

Acknowledgements

Jennifer Meyer is supported by a Jacobs Foundation Research Fellowship (2024–2026).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.compedu.2025.105386>.

Data availability

Data used in this study are available on OSF for independent validation of the results. The materials can be found at: <https://osf.io/e6vuy/>.

References

- Alves, R. A., Castro, S. L., & Olive, T. (2008). Execution and pauses in writing narratives: Processing time, cognitive effort and typing skill. *International Journal of Psychology*, 43(6), 969–979. <https://doi.org/10.1080/00207590701398951>
- Bahr, J. L., Höft, L., Lipnevich, A., Meyer, J., & Jansen, T. (2025). Exploring students' receptivity to feedback: A latent profile analysis. *Assessment in Education: Principles, Policy & Practice*, 1–19. <https://doi.org/10.1080/0969594X.2025.2467676>
- Bennett, R. E., & Zhang, M. (2015). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing* (pp. 142–173). Routledge. <https://doi.org/10.4324/9781315871493-8>
- Biedermann, D., Ciordas-Hertel, G.-P., Winter, M., Mordel, J., & Drachsler, H. (2023). Contextualized logging of on-task and off-task behaviours during learning. *Journal of Learning Analytics*, 10(2), 115–125. <https://doi.org/10.18608/jla.2023.7837>
- Bolzer, M., Strijbos, J.-W., & Fischer, F. (2015). Inferring mindful cognitive-processing of peer-feedback via eye-tracking: Role of feedback-characteristics, fixation-durations and transitions. *Journal of Computer Assisted Learning*, 31(5), 422–434. <https://doi.org/10.1111/jcal.12091>
- Bouwer, R., & Dirks, K. (2023). The eye-mind of processing written feedback: Unraveling how students read and use feedback for revision. *Learning and Instruction*, 85, Article 101745. <https://doi.org/10.1016/j.learninstruc.2023.101745>
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, 26(2), 211–252. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>
- Bråten, I., Latini, N., & Haverkamp, Y. E. (2022). Predictors and outcomes of behavioral engagement in the context of text comprehension: When quantity means quality. *Reading and Writing*, 35(3), 687–711. <https://doi.org/10.1007/s11145-021-10205-x>
- Chan, S., Lo, N., & Wong, A. (2024). Generative AI and essay writing: Impacts of automated feedback on revision performance and engagement. *reFlections*, 31(3), 1249–1284. <https://doi.org/10.61508/refl.v31i3.277514>
- Choi, H., Winne, P. H., Brooks, C., Li, W., & Shedden, K. (2023). Logs or self-reports? Misalignment between behavioral trace data and surveys when modeling learner achievement goal orientation. *Proceedings of the 13th international learning analytics and knowledge conference (LAK)*. <https://doi.org/10.1145/3576050.3576052>
- Conijn, R., Cook, C., van Zaanen, M., & Van Waes, L. (2021). Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 32(4), 835–866. <https://doi.org/10.1007/s40593-021-00268-w>
- Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Benner, M., & Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, 54, Article 100667. <https://doi.org/10.1016/j.asw.2022.100667>
- de Smet, M. J. R., Leijten, M., & Van Waes, L. (2018). Exploring the process of reading during writing using eye tracking and keystroke logging. *Written Communication*, 35(4), 411–447. <https://doi.org/10.1177/0741088318788070>
- Dux Speltz, E., & Chukharev-Hudilainen, E. (2021). The effect of automated fluency-focused feedback on text production. *Journal of Writing Research*, 13(2), 231–255. <https://doi.org/10.17239/jowr-2021.13.02.02>
- Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition & Communication*, 32(4), 400. <https://doi.org/10.2307/356602>
- Fan, Y., Rakovic, M., Van Der Graaf, J., Lim, L., Singh, S., Moore, J., Molenaar, I., Bannert, M., & Gašević, D. (2023). Towards a fuller picture: Triangulation and integration of the measurement of self-regulated learning based on trace and think aloud data. *Journal of Computer Assisted Learning*, 39(4), 1303–1324. <https://doi.org/10.1111/jcal.12801>
- Fleckenstein, J., Jansen, T., Meyer, J., Trüb, R., Raubach, E. E., & Keller, S. D. (2024). How am I going? Behavioral engagement mediates the effect of individual feedback on writing performance. *Learning and Instruction*, 93, Article 101977. <https://doi.org/10.1016/j.learninstruc.2024.101977>
- Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6, Article 1162454. <https://doi.org/10.3389/frai.2023.1162454>
- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition & Communication*, 32(4), 365. <https://doi.org/10.2307/356600>
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59–109. <https://doi.org/10.3102/00346543074001059>
- Garcés-Manzanera, A. (2023). Young EFL learners' pausing behavior: Exploring pause thresholds in two proficiency levels. *ESPIRAL. CUADERNOS DEL PROFESORADO*, 16(32), 64–75. <https://doi.org/10.25115/ecp.v16i32.9047>

- Gašević, D., Jovanovic, J., Pardo, A., & Dawson, S. (2017). Detecting learning strategies with analytics: Links with self-reported measures and academic performance. *Journal of Learning Analytics*, 4(2). <https://doi.org/10.18608/jla.2017.42.10>
- Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: On validating the interpretation of process indicators based on log data. *Large-Scale Assessments in Education*, 9(1), 20. <https://doi.org/10.1186/s40536-021-00113-5>
- Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal*, 115(4), 523–547. <https://doi.org/10.1086/681947>
- Greene, B. A., & Miller, R. B. (1996). Influences on achievement: Goals, perceived ability, and cognitive engagement. *Contemporary Educational Psychology*, 21(2), 181–192. <https://doi.org/10.1006/ceps.1996.0015>
- Guo, H., Deane, P. D., Van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, 55(2), 194–216. <https://doi.org/10.1111/jedm.12172>
- Hall, S., Baaijen, V. M., & Galbraith, D. (2024). Constructing theoretically informed measures of pause duration in experimentally manipulated writing. *Reading and Writing*, 37(2), 329–357. <https://doi.org/10.1007/s11145-022-10284-4>
- Han, Y., & Hyland, F. (2019). Learner engagement with written feedback: A sociocognitive perspective. In K. Hyland, & F. Hyland (Eds.), *Feedback in second language writing* (2nd ed., pp. 247–264). Cambridge University Press. <https://doi.org/10.1017/9781108635547.015>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Publications.
- Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education*, 90, 36–53. <https://doi.org/10.1016/j.compedu.2015.09.005>
- Horbach, A., Laarmann-Quante, R., Liebenow, L., Jansen, T., Keller, S., Meyer, J., Zesch, T., & Fleckenstein, J. (2022). Bringing automatic scoring into the classroom – Measuring the impact of automated analytic feedback on student writing performance. *Proceedings of the 11th workshop on natural language processing for computer-assisted language learning (NLP4CALL 2022)*. <https://doi.org/10.3384/ecp190008>
- Jansen, T., Höft, L., Bahr, L., Fleckenstein, J., Möller, J., Köller, O., & Meyer, J. (2024). Comparing generative AI and expert feedback to students' writing: Insights from student teachers. *Psychologie in Erziehung und Unterricht*, 71(2), 80–92. <https://doi.org/10.2378/peu2024.art08d>
- Järvelä, S., Veermans, M., & Leinonen, P. (2008). Investigating student engagement in computer-supported inquiry: A process-oriented analysis. *Social Psychology of Education*, 11(3), 299–322. <https://doi.org/10.1007/s11218-007-9047-6>
- Keller, S. D., Fleckenstein, J., Krüger, M., Köller, O., & Rupp, A. A. (2020). English writing skills of students in upper secondary education: Results from an empirical study in Switzerland and Germany. *Journal of Second Language Writing*, 48, Article 100700. <https://doi.org/10.1016/j.jslw.2019.100700>
- Keller, S. D., Trüb, R., Raubach, E., Meyer, J., Jansen, T., & Fleckenstein, J. (2023). Designing and validating an assessment rubric for writing emails in English as a foreign language. *Research in Subject-Matter Teaching and Learning (RISTAL)*, 6(1), 16–48. <https://doi.org/10.2478/ristal-2023-0002>
- Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & Education*, 104, 18–33. <https://doi.org/10.1016/j.compedu.2016.10.001>
- Kovanovic, V., Gašević, D., Dawson, S., Joksimovic, S., & Baker, R. (2016). Does time-on-task estimation matter? Implications on validity of learning analytics findings. *Journal of Learning Analytics*, 2(3), 81–110. <https://doi.org/10.18608/jla.2015.23.6>
- Lacruz, I., Shreve, G. M., & Angelone, E. (2012). Average pause ratio as an indicator of cognitive effort in post-editing: A case study. *Association for machine translation in the americas, workshop on post-editing technology and practice*. <https://aclanthology.org/2012.amta-wptp.3>
- Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using inputlog to analyze and visualize writing processes. *Written Communication*, 30(3), 358–392. <https://doi.org/10.1177/0741088313491692>
- Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Cybern Control Theory*, 10(8), 708–710.
- Lindgren, E., & Sullivan, K. (2006). Writing and the analysis of revision: An overview. In K. Sullivan, & E. Lindgren (Eds.), *Computer key-stroke logging and writing* (pp. 31–44). BRILL. https://doi.org/10.1163/9780080460932_004
- Lipnevich, A. A., & Smith, J. K. (2022). Student – Feedback interaction model: Revised. *Studies In Educational Evaluation*, 75, Article 101208. <https://doi.org/10.1016/j.stueduc.2022.101208>
- Martin, A. J. (2023). Integrating motivation and instruction: Towards a unified approach in educational psychology. *Educational Psychology Review*, 35(2), 54. <https://doi.org/10.1007/s10648-023-09774-w>
- Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, 30(6), 1267–1285. <https://doi.org/10.1007/s11145-017-9723-7>
- Mertens, U., Finn, B., & Lindner, M. A. (2022). Effects of computer-based feedback on lower- and higher-order learning outcomes: A network meta-analysis. *Journal of Educational Psychology*, 114(8), 1743–1772. <https://doi.org/10.1037/edu0000764>
- Meyer, J., Jansen, T., & Fleckenstein, J. (2025). Nonengagement and unsuccessful engagement with feedback in lower secondary education: The role of student characteristics. *Contemporary Educational Psychology*, 81, Article 102363. <https://doi.org/10.1016/j.cedpsych.2025.102363>
- Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6, Article 100199. <https://doi.org/10.1016/j.caeai.2023.100199>
- Miller, B. W. (2015). Using reading times and eye-movements to measure cognitive engagement. *Educational Psychologist*, 50(1), 31–42. <https://doi.org/10.1080/00461520.2015.1004068>
- Ministerium für Bildung und Wissenschaft des Landes Schleswig-Holstein [Ministry of Education and Science Schleswig-Holstein]. (2014). <https://fachportal.lernnetz.de/sh/faecher/englisch/fachanforderungen.html>
- Mohsen, M. A., & Qassem, M. (2020). Analyses of L2 learners' text writing strategy: Process-oriented perspective. *Journal of Psycholinguistic Research*, 49(3), 435–451. <https://doi.org/10.1007/s10936-020-09693-9>
- Narciss, S. (2008). Feedback strategies for interactive learning tasks. In D. Jonassen, M. J. Spector, M. Driscoll, M. D. Merrill, J. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 125–143). Routledge.
- 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. <https://doi.org/10.1080/10494820.2022.2096642>
- Nguyen, T. T. D. T., Garncaz, T., Ng, F., Dabbish, L. A., & Dow, S. P. (2017). Fruitful feedback: Positive affective language and source anonymity improve critique reception and work outcomes. *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. <https://doi.org/10.1145/2998181.2998319>
- OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD. <https://doi.org/10.1787/dfe0bf9c-en>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. <https://doi.org/10.1126/science.aac4716>
- Osborne, J. (2010). Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research and Evaluation*, 15(12). <https://doi.org/10.7275/QBPC-GK17>
- Panadero, E. (2023). Toward a paradigm shift in feedback research: Five further steps influenced by self-regulated learning theory. *Educational Psychologist*, 58(3), 193–204. <https://doi.org/10.1080/00461520.2023.2223642>
- Panadero, E., & Lipnevich, A. (2022). A review of feedback typologies and models: Towards an integrative model of feedback elements. *Educational Research Review*, 35, Article 100416. <https://doi.org/10.1016/j.edurev.2021.100416>
- Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLoS One*, 19(5), Article e0304013. <https://doi.org/10.1371/journal.pone.0304013>
- Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, 16(2), 93–115. <https://doi.org/10.1037/a0022658>

- Price, M., Handley, K., & Millar, J. (2011). Feedback: Focusing attention on engagement. *Studies in Higher Education*, 36(8), 879–896. <https://doi.org/10.1080/03075079.2010.483513>
- Quick, J., Motz, B., Israel, J., & Kaetzel, J. (2020). What college students say, and what they do: Aligning self-regulated learning theory with behavioral logs. *Proceedings of the tenth international conference on learning analytics & knowledge (LAK)*. <https://doi.org/10.1145/3375462.3375516>
- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing [Computer software] Version 4.4.1. <https://www.R-project.org/>.
- Reinhold, F., Leuders, T., Loibl, K., Nückles, M., Beege, M., & Boelmann, J. M. (2024). Learning mechanisms explaining learning with digital tools in educational settings: A cognitive process framework. *Educational Psychology Review*, 36(1), 14. <https://doi.org/10.1007/s10648-024-09845-6>
- Révész, A., Kourtali, N., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning*, 67(1), 208–241. <https://doi.org/10.1111/lang.12205>
- Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed-methods study. *Studies in Second Language Acquisition*, 41(3), 605–631. <https://doi.org/10.1017/S027226311900024X>
- Roeser, J., Conijn, R., Chukharev, E., Ofstad, G. H., & Torrance, M. (2025). Typing in tandem: Language planning in multisentence text production is fundamentally parallel. *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001759>
- Roeser, J., De Maeyer, S., Leijten, M., & Van Waes, L. (2024). Modelling typing disfluencies as finite mixture process. *Reading and Writing*, 37(2), 359–384. <https://doi.org/10.1007/s11145-021-10203-z>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2). <https://doi.org/10.18637/jss.v048.i02>
- Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: A case study in Switzerland and Germany. *ETS Research Report Series*, 2019(1), 1–23. <https://doi.org/10.1002/ets2.12249>
- Sailer, M., Ninaus, M., Huber, S. E., Bauer, E., & Greiff, S. (2024). The end is the beginning is the end: The closed-loop learning analytics framework. *Computers in Human Behavior*, 158, Article 108305. <https://doi.org/10.1016/j.chb.2024.108305>
- Schiller, R., Fleckenstein, J., Mertens, U., Horbach, A., & Meyer, J. (2024). Understanding the effectiveness of automated feedback: Using process data to uncover the role of behavioral engagement. *Computers & Education*, 223, 105163. <https://doi.org/10.1016/j.compedu.2024.105163>
- Schuurman, A. R., Baarsma, M. E., Wiersinga, W. J., & Hovius, J. W. (2022). Digital disparities among healthcare workers in typing speed between generations, genders, and medical specialties: Cross sectional study. *BMJ*, Article e072784. <https://doi.org/10.1136/bmj-2022-072784>
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76. <https://doi.org/10.1016/j.asw.2013.04.001>
- Smiley, W., & Anderson, R. (2011). Measuring students' cognitive engagement on assessment tests: A confirmatory factor analysis of the short form of the cognitive engagement scale. *Research & Practice in Assessment*, 6, 17–28.
- Steiss, J., Tate, T. P., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91, Article 101894. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Talebinamvar, M., & Zarrabi, F. (2022). Clustering students' writing behaviors using keystroke logging: A learning analytic approach in EFL writing. *Language Testing in Asia*, 12(1), 6. <https://doi.org/10.1186/s40468-021-00150-5>
- Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5). <https://doi.org/10.18637/jss.v059.i05>
- Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes. *Review of Educational Research*, 85(4), 475–511. <https://doi.org/10.3102/0034654314564881>
- Van Waes, L., van Weijen, D., & Leijten, M. (2014). Learning to write in an online writing center: The effect of learning styles on the writing process. *Computers & Education*, 73, 60–71. <https://doi.org/10.1016/j.compedu.2013.12.009>
- Vandermeulen, N., Leijten, M., & 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), 109–140. <https://doi.org/10.17239/jowr-2020.12.01.05>
- Vandermeulen, N., Lindgren, E., Waldmann, C., & Levlin, M. (2024). Getting a grip on the writing process: (effective) approaches to write argumentative and narrative texts in L1 and L2. *Journal of Second Language Writing*, 65, Article 101113. <https://doi.org/10.1016/j.jslw.2024.101113>
- Vandermeulen, N., Van Steendam, E., De Maeyer, S., & Rijlaarsdam, G. (2023). Writing process feedback based on keystroke logging and comparison with exemplars: Effects on the quality and process of synthesis texts. *Written Communication*, 40(1), 90–144. <https://doi.org/10.1177/07410883221127998>
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics* with S. New York: Springer. <https://doi.org/10.1007/978-0-387-21706-2>
- Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94–109. <https://doi.org/10.1016/j.compedu.2016.05.004>
- Winstone, N. E., & Nash, R. A. (2023). Toward a cohesive psychological science of effective feedback. *Educational Psychologist*, 58(3), 111–129. <https://doi.org/10.1080/00461520.2023.2224444>
- Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of reciprocity processes. *Educational Psychologist*, 52(1), 17–37. <https://doi.org/10.1080/00461520.2016.1207538>
- Winter, M., Mordel, J., Mendzheritskaya, J., Biedermann, D., Ciordas-Hertel, G.-P., Hahnel, C., Bengs, D., Wolter, I., Goldhammer, F., Drachler, H., Artelt, C., & Horz, H. (2024). Behavioral trace data in an online learning environment as indicators of learning engagement in university students. *Frontiers in Psychology*, 15, Article 1396881. <https://doi.org/10.3389/fpsyg.2024.1396881>
- Wong, Z. Y., & Liem, G. A. D. (2022). Student engagement: Current state of the construct, conceptual refinement, and future research directions. *Educational Psychology Review*, 34(1), 107–138. <https://doi.org/10.1007/s10648-021-09628-3>
- Wong, Z. Y., Liem, G. A. D., Chan, M., & Datu, J. A. D. (2024). Student engagement and its association with academic achievement and subjective well-being: A systematic review and meta-analysis. *Journal of Educational Psychology*, 116(1), 48–75. <https://doi.org/10.1037/edu0000833>
- Xu, C. (2024). The effects of topic familiarity on L2 writing—A process-product approach. <https://doi.org/10.2139/ssrn.4903705>
- Zesch, T., & Horbach, A. (2018). ESCRITO – An nlp-enhanced educational scoring toolkit. *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.
- Zhai, N., & Ma, X. (2022). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, 61(4), 875–900. <https://doi.org/10.1177/07356331221127300>
- Zhang, M., & Deane, P. (2015). Process features in writing: Internal structure and incremental value over product features. *ETS Research Report Series*, 2015(2), 1–12. <https://doi.org/10.1002/ets2.1207>
- Zhang, M., Zhu, M., Deane, P., & Guo, H. (2019). Identifying and comparing writing process patterns using keystroke logs. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (pp. 367–381). Springer International Publishing. https://doi.org/10.1007/978-3-030-01310-3_32
- Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, 37(2), 197–206. <https://doi.org/10.1086/651257>
- Zhu, M., Liu, O. L., & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education*, 143, Article 103668. <https://doi.org/10.1016/j.compedu.2019.103668>