International Journal of Artificial Intelligence in Education (2022) 32:835–866
https://doi.org/10.1007/s40593-021-00268-w

ARTICLE

Early prediction of writing quality using keystroke logging

Rianne Conijn[1] · Christine Cook[2] · Menno van Zaanen[3] · Luuk Van Waes[4]

Accepted: 10 July 2021 / Published online: 24 August 2021

Abstract

Feedback is important to improve writing quality; however, to provide timely and personalized feedback is a time-intensive task. Currently, most literature focuses on providing (human or machine) support on product characteristics, especially after a draft is submitted. However, this does not assist students who struggle *during* the writing process. Therefore, in this study, we investigate the use of keystroke analysis to predict writing quality throughout the writing process. Keystroke data were analyzed from 126 English as a second language learners performing a timed academic summarization task. Writing quality was measured using participants' final grade. Based on previous literature, 54 keystroke features were extracted. Correlational analyses were conducted to identify the relationship between keystroke features and writing quality. Next, machine learning models (regression and classification) were used to predict final grade and classify students who might need support at several points during the writing process. The results show that, in contrast to previous work, the relationship between writing quality and keystroke data was rather limited. None of the regression models outperformed the baseline, and the classification models were only slightly better than the majority class baseline (highest AUC = 0.57). In addition, the relationship between keystroke features and writing quality changed throughout the course of the writing process. To conclude, the relationship between keystroke data and writing quality might be less clear than previously posited.

Keywords Keystroke logging · Early prediction · Writing quality · Academic writing · Writing processes

✉ Rianne Conijn
m.a.conijn@tue.nl

[1] Human-Technology Interaction Group, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands

[2] Department of Communication, National Chengchi University, Taipei, Taiwan

[3] South African Centre of Digital Language Resources, Potchefstroom, South Africa

[4] Department of Management, University of Antwerp, Antwerp, Belgium

Introduction

Feedback on writing plays a key role in improving writing quality and writing proficiency (Bitchener et al., 2005; Chandler, 2003; Graham et al., 2015; Parr & Timperley, 2010). For feedback to be effective, it needs to be timely and frequent (Cotos, 2015; Ferguson, 2011; Shute, 2008). However, providing timely feedback on academic writing in higher education is complex and time-consuming. Accordingly, several automated feedback systems have been developed to augment teacher feedback and consequently enhance students' writing proficiency (Dikli, 2006; Passonneau et al., 2017; Stevenson & Phakiti, 2014).

A key part of most of these systems is the automated scoring of students' writing, as this measure of writing quality can indicate which students are in need of support. Automated scoring systems have demonstrated the ability to provide fairly accurate predictions of human scores based on properties of the writing product, e.g., number of words in a draft (Allen et al., 2015). In this way, a score is predicted once a writing product (draft or final version) is finished. However, for *timely* feedback, we would like to be able to predict writing scores as soon as possible in order to identify the students who would need support or are 'at risk' of failing the assignment *before* the writing product is finished (Romero & Ventura, 2019).

For the early prediction of writing scores, i.e., prediction of writing scores before the draft has been finished, there are two possible approaches. First, we can predict writing scores based on snapshots of the text produced so far. However, the text produced so far might not contain enough information (e.g., limited number of words) to provide accurate predictions. Second, we can predict writing scores based on information on the writing process, such as the number of revisions made. This approach has the additional advantage that it may be used to provide feedback both *during* and *on* the writing process. Feedback on the writing process is more powerful for deep processing than feedback on the product, as it can not only be used to improve the current task, but also to enhance (process) skills that can be transferred to future tasks (Hattie & Timperley, 2007; Vandermeulen et al., 2020).

Several studies already show that information on the writing process can be used to accurately predict holistic grades (e.g., Allen et al., 2016; Choi & Deane, 2021; Guo et al., 2018; Sinharay et al., 2019). In these studies, keystroke logging was used to extract information on the students' writing processes (Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). However, these studies used a varying set of keystroke features, making it hard to determine which features are most relevant for predicting writing quality. In addition, these studies only analyzed the writing process *after* it was finished, using the full keystroke log. Hence, it is still unknown whether it is possible to use information on the writing process for the early prediction of writing quality.

Therefore, in this study, we aim to further explore the demonstrated relationship between keystroke data and writing quality of English as a second language (ESL) students. Specifically, we aim to answer the following research questions:

RQ1) Which keystroke features identified in previous literature are related to writing quality?
RQ2) Can keystroke features identified in previous literature be used for timely writing quality prediction at different stages in the writing process?
RQ3) Which keystroke features are important for timely writing quality prediction at different stages in the writing process?

To answer these research questions, we operationalized writing quality as the final score of ESL students writing an academic summary. Keystroke analysis is used to automatically extract information on students' writing processes. Of course, we should keep in mind that keystroke logging data are always the result of indirect observation, and we should therefore be cautious when interpreting the relationship between keystroke features and the underlying (cognitive) subprocesses (Galbraith & Baaijen, 2019).

Timing of writing feedback

Human feedback on writing is almost universally provided after the completion of a draft or final version (Gielen et al., 2010; Graham et al., 2015; Parr & Timperley, 2010). This feedback timing can be explained by the fact that the assessment and resulting feedback is commonly aimed at drafts or (near) finalized products. Teachers provide little feedback during writing, as this requires insight into the writing process, which is primarily obtained via time-intensive methods like thinking-aloud procedures and observational studies (see e.g., Beauvais et al., 2011; Braaksma et al., 2004).

While human feedback on the writing product usually only occurs once or twice per writing assignment, computer-based support for writing allows for timely assessment and feedback on multiple drafts (Cotos, 2015). There is a large variety of computer-based writing support systems available to assist teachers in providing writing support, such as automated essay scoring (AES), automated writing evaluation (AWE), and intelligent tutoring systems (ITS; Allen et al., 2015). AES are grading systems typically used for summative assessment that can be used as either an alternative to teachers' grading or as a first-draft evaluator (Dikli, 2006; Wilson, 2017). AWE systems are intended as formative assessment tools, providing more detailed feedback and suggestions for improvement than AES systems (Cotos, 2015). Lastly, ITS extend on AWE systems by also providing instructional content, probing questions and interactivity (Ma et al., 2014). These systems can provide feedback on multiple drafts, but also often offer the possibility for dynamic feedback during the writing process or feedback on demand. These systems provide feedback on different aspects of the writing product, such as spelling and grammar (e.g. Grammarly, see Koltovskaia, 2020), rhetorical moves (e.g., AcaWriter, see Knight et al., 2020), or even on specific parts of an academic thesis (e.g., Thesis Writer, see Rapp & Kauf, 2018).

However, just like with human support, the automated assessment and feedback is still commonly based on the writing product (or intermediate writing products), rather than the writing process (Cotos, 2015; Ma et al., 2014; Wang et al., 2013). Accordingly, the feedback commonly targets revisions at the micro (product) level, such as grammar and wording, rather than support for the development of writing strategies and self-monitoring (Strobl et al., 2019). Therefore, in this article, we focus on the writing process; for an overview of automatic feedback on the writing product, see Crossley (2020) and Dikli (2006). To provide feedback on the writing process, we first need to identify the students who are in need of support.

Measuring writing processes with keystroke data

Keystroke analysis has been increasingly used to gain insight into the writing process. With keystroke logging, information on every key pressed and released is recorded, resulting in a detailed log on students' text composition (Leijten & Van Waes, 2013; Lindgren & Sullivan, 2019). Keystroke logging observation has advantages compared to other methods of collecting data about the writing process. First, keystroke logging can provide objective, detailed, and real-time information on students' unfolding typing processes during their writing. Second, keystroke logging is less intrusive and more scalable than traditional observational methods, such as screen recording and thinking-aloud.

Given the fine-grained nature of keystroke logging data, a variety of features have been extracted for keystroke analysis. Based on previous literature, we distinguish five different groups of keystroke features: (1) features related to pause timings, such as interkeystroke intervals or timings between words (Barkaoui, 2016; Medimorec & Risko, 2017); (2) features related to revisions, such as the number of backspaces or the duration of backspacing events (Barkaoui, 2016; Deane, 2014); (3) features related to verbosity, such as the number of words (Allen et al., 2016; Likens et al., 2017); (4) features related to fluency, such as the percentage of bursts ending in a revision (Baaijen et al., 2012; Van Waes & Leijten, 2015); and (5) features related to events other than keystrokes producing characters, such as text selections, insertions (paste), deletions (cut), and mouse movements (Baaijen & Galbraith, 2018; Leijten et al., 2019).

The feature selection and analysis of keystroke data largely depend on the methodological approach, which can be roughly divided into theory-driven and data-driven approaches. Theory-driven approaches aim to connect behavioral keystroke data with cognitive writing processes. These studies use a select set of keystroke features, frequently triangulated with other data such as manual annotations or think-aloud data, within a tightly-controlled experimental setting. For example, several studies link keystroke data to the three writing processes as defined by Flower and Hayes (1980): planning, translating, and reviewing processes (see e.g., Galbraith & Baaijen, 2019; Tillema et al., 2011). For example, pauses between words have been seen as an indicator of planning, retrieving, and editing processes (Baaijen et al., 2012; Medimorec & Risko, 2017), and the number of deletions have been used as an indicator of writing fluency and revision processes (Van Waes et al., 2014). This shows that keystroke data, at least to some extent, can be used as a proxy for writing processes.

By contrast, data-driven studies often include as many features as necessary to build a scalable and accurate model for the problem at hand (see e.g., Bixler & D'Mello, 2013). In the current study, a data-driven approach is used to automatically predict writing quality during the writing process. Hence, we cannot (and do not intend to) make any theoretical claims about the relation between cognitive writing processes and behavioral keystroke data. Rather, we aim to identify whether keystroke data, which provides information on the writing process, relates to writing quality.

Relation between keystroke data and writing quality

Before information on the writing process was measured using keystroke logs, several relationships were found between higher-level writing processes, such as planning and revision, and writing quality. In terms of planning, Kellogg (1987) found that preparing a written outline enhances writing quality, but that drafting style (rough versus polished) did not have an effect on writing quality. However, other studies did find an effect, where detailed plans resulted in higher writing quality than minimal drafts (Torrance et al., 2000). In terms of revisions, more proficient writers revise more and focus more on meaning level revisions than less proficient writers who focus more on surface level revisions, such as punctuation, spelling, and wording (Choi, 2007; Faigley & Witte, 1981). The influence of these writing processes on writing quality differs over time. For example, reading the assignment and evaluating the text written so far are positively related to quality in the beginning, but negatively in the middle of the writing process. Likewise, goal setting, generating ideas, structuring, rereading, and writing are positively related to writing quality at the end of the writing process, but either negatively or unrelated in the beginning (Breetvelt et al., 1994).

With the advent of keystroke logging in writing research, more fine-grained measures of specific aspects of the writing process have been related to writing quality across a variety of tasks, such as argumentative and policy recommendation essays (Guo et al., 2018; Zhang et al., 2016), as well as persuasive essays (Allen et al., 2016; Deane, 2014; Likens et al., 2017; Sinharay et al., 2019; Zhang et al., 2016). First, total time on task has been shown to be positively correlated with writing scores in several studies, with correlations ranging from 0.40 to 0.52 (Guo et al., 2018; Sinharay et al., 2019; Zhang et al., 2019). Features related to pause timings, such as pauses within words, have been found to be negatively related with writing scores (r = -0.36; Sinharay et al., 2019), while vectors of interword intervals have been found to be positively related (r = 0.46 to 0.48, Zhang et al., 2016). In addition, for revisions, students with low second language proficiency made more revisions, and especially more typographic, language, and pre-contextual revisions (revisions at the leading edge), compared to students with high second language proficiency (Barkaoui, 2016; Xu, 2018). Moreover, features related to verbosity, such as the number of keystrokes (r = 0.59; Zhang et al., 2019) or the number of words (r = 0.53, Likens et al., 2017) are positively related to writing scores. Lastly, features related to writing fluency, such as the typing speed ($r = 0.31$ to 0.39; Sinharay et al., 2019; Zhang et al., 2019), number of bursts (sequences of keystrokes without a long pause; $r = 0.49$), and burst length ($r = 0.38$), have also been found to be positively related to writing quality (Sinharay et al., 2019).

In addition to the correlational analyses with single features, multiple features have also been combined to predict writing quality. For example, Allen et al. (2016) showed that the number of words, number of backspaces, maximum and median interkeystroke interval, standard deviation, entropy, and maximum of the number of keystrokes per 30-s interval, and the standard deviation of the distance between 30-s windows with at least one keystroke, could explain 74% of the variance in essay scores. The number of keystrokes was found to be most predictive of essay score. A later study using the first 999 keystrokes of the same dataset showed that 28% of the variance in essay score could be explained by the number of words (Likens et al., 2017). Furthermore, fractal properties from multifractal analysis on the interkeystroke interval timeseries, combined with the number of words, could explain 35% of the variance in the essay score. Using boosting with regression trees, Sinharay et al. (2019) were able to predict writing scores with 38 process features, leading to an RMSE of 0.50 (on a scale from 1–5). Time on task, typing speed, number of bursts, and burst length had the most predictive power. The prediction with product features was only slightly better (RMSE = 0.44) than the process features. Adding process features to the product features did not enhance the prediction accuracy. Similarly, Choi and Deane (2021) showed that writing quality of two source-based writing tasks and 2 argumentative essays could be explained by only two to five features, leading to a MSE of 0.49–0.86 (on a scale from 1–5). The number of keystrokes was the most important predictor, and the only predictor that proved to be important across all tasks.

Lastly, some studies used feature reduction on the keystroke features prior to the prediction of writing quality. For example, Deane (2014) identified three factors in the logged data: latency, editing behavior, and burst span. Combined, these factors could explain 60% of the variance in essay scores for the persuasive task, and 68% of the variance for the literacy analysis task. Another study identified two factors: sentence production and global linearity (Baaijen & Galbraith, 2018). These factors, however, did not show significant correlations with text quality.

Given the differences in sample sizes, grading procedure, writing task, writing environment, keystroke features extracted, and analyses used, the results of these studies are hard to compare. That said, they *do* provide some insight into which keystroke features are related to writing quality, and to what extent writing quality can be predicted using information from the logged data. However, these studies used different and relatively small sets of keystroke features, making it hard to determine which of the features are most relevant for predicting writing quality. In addition, the keystroke features in the reported studies were analyzed after the writing process was finished. It is still unknown at which stages in the writing process keystroke data can be used for the timely prediction of writing quality.

Timely identification of students who need support

Timely identification of students who might need support (sometimes referred to as students ‘at risk’) is a common theme in the fields of learning analytics and educational data mining (Romero & Ventura, 2019). It has been shown that these students could be identified relatively quickly in a variety of contexts and with a variety of datasets. For example, at the course level, learning management system data can be used to identify students who might need support early on in the course (Macfadyen & Dawson, 2010), and historical data on grades and courses taken can be used to determine which students might need support even before the course has started (Polyzou & Karypis, 2019). At the task level, prior performance, hint usage, activity progress, and interface interaction can be used to predict successful completion of block-based programming tasks (Emerson et al., 2019), and clickstream data can be used to predict the successful completion of a novice programming task within the first minute of the task (Mao et al., 2019).

To the authors’ knowledge, no studies have looked into the early prediction of which students might need support using keystroke data, with the exception of Casey’s (2017) study using keystroke data to predict performance in a programming course after each week in the semester. However, no keystroke studies have looked into the early identification of students who might need support during (academic) writing.

Current study

In the current study, we aim to determine the relationship between the writing process (measured by keystroke data) and writing quality (measured by final grade). To determine which keystroke features (obtained from previous work) are related to writing quality (RQ1), we first conduct correlational analyses. Thereafter, two types of prediction algorithms are trained to determine which keystroke features can be used for the early prediction of writing quality and how prediction accuracy changes over time (RQ2). Specifically, machine learning algorithms are trained to predict writing scores (regression) and to predict which students might need support (binary classification) at different stages in the writing process. Lastly, feature importance is calculated for all features in all the models to determine which keystroke features are important at which stages in the writing process (RQ3).

Method

Participants

The data used in this study were collected during an academic writing course for ESL learners. This course was taught for premaster students of communication and information sciences. The study consisted of an online part and an in-lecture part. In total, 141 students participated in the online part, and 130 students participated in both parts (nine participants did not show, two participants did not provide consent). Of these 130 participants, two participants did not type in the specified Word document, one participant only completed the copy task, and one participant only copy-pasted text. Therefore, these participants were excluded, resulting in a total of 126 participants remaining for analysis. Of the 126 participants, 84 (67%) were female and the average age was 24 ($SD = 2.7$). Most participants were native Dutch speakers (95 participants, 75%), and for the large majority, this was their first academic writing course (118 participants, 94%).

Procedure

In the online part, participants were asked to provide informed consent and to complete a questionnaire on demographics and self-reported writing style. Following their regular class procedure, they were asked to read a given English-language journal article in preparation for the lecture the following week. The article involved a 2×2 experimental design setup in the field of their premaster program (Woong Yun & Park, 2011). A week after the questionnaire, the participants were assigned two writing tasks during the lecture. The first task was a copy task in which they had to transcribe a given fable of 850 characters. The second task was an academic summary task, where the participants were asked to write an academic summary of 100–200 words based on the article they read in the week prior to the lecture (the abstract was removed from the article). The students were familiar with this task, as they were also asked to perform a similar task on a different article earlier in the course. The participants were allowed 30 min to finish this task. Five minutes before the end of the task, they were reminded to finish their writing. Upon completion of the task, participants were asked to provide a second consent, in which they were asked to indicate they did not type any personal information during the task and still agreed to their data being used.

All students used similar desktop computers for the task. The task description was shown on a single page at the left of the screen, while the Word document where participants could type the summary was on the right. The text of the journal article was added underneath the task description in the same document. To consult the journal article, participants had to focus (click) on the task description and scroll down. The participants were allowed to use the Internet during the task (for example, to consult an online dictionary). During the tasks, keystroke data and mouse data were collected using Inputlog (Leijten & Van Waes, 2013).

Data collection

For the current study, the keystroke data of the academic writing task and the essay scores were analyzed. The keystroke data of the copy task were used as a reference task for extracting the keystroke features, so as to obtain an average baseline pause time and interkeystroke interval (IKI) for each participant.

The academic writing tasks were independently graded by two of the authors, one native English speaker and one ESL speaker, both highly experienced in grading writing. The writing tasks were scored against five rubrics: main idea, structure and organization, content, language and paraphrasing, and grammar and mechanics from 1 (not passing) to 4 (exceptional; see Appendix A). This grading scheme was known to the students and was also used in the previous, similar academic summary task in the course. The final grade was calculated by summing the rubric grades and dividing them by 2. Points were subtracted if the student did not comply with the task (e.g., wrote too few or too many words, or did not cite the authors). This resulted in a scale from 1 to 10. As a final grade of 5.5 or higher is considered a pass in the grading system used at the university, a student was considered who might need support if the predicted final grade was lower than 5.5.

The Inter-rater reliability of the final grade was calculated using a quadratically weighted kappa to account for the ordinal scale (as in Zhang et al., 2019). There was a strong correlation between the grades of the two raters ($r=0.88$) and the inter-rater reliability was high ($\kappa=0.89$). As in Zhang et al. (2019), the grades from the first grader (native English speaker), were used for analysis. On average, the students scored 5.50 ($SD=1.36$) and 60 students scored below 5.5, indicating they might have needed support (48%).

Feature extraction

Based on existing literature, a total of 54 features were extracted from the keystroke log. Extant literature sometimes uses different definitions for similar keystroke features, and it is sometimes unclear how exactly a certain feature is extracted. For replicability, we provide a detailed overview of the extraction of features from the keystroke log. To be able to replicate the features from previous studies as closely as possible, we did not use the analysis tool provided by Inputlog, but rather the raw data (’basic log file’) to extract the features in R. The R scripts for the feature extraction, feature reduction, and model building can be accessed at https://github.com/anonymized.

Inputlog collects both keystroke and mouse data, and distinguishes five types of events: keyboard, mouse, insert (insertion of text from within the document or other source), replacement (selection of text), and focus (click on another window, e.g., another document or web page). Here, we define a keystroke as a keyboard event. This includes any key pressed and includes character key presses (e.g., ’a’, ’6’, or ’$’), as well as control, function, or navigation keys (e.g., ’Alt’, ’F5’, ’Home’). Sometimes multiple keystrokes are required to generate one character (e.g., for capitalization). For every character typed, the location in the document where it is typed is stored. We distinguished two locations: leading edge (at most two characters away from the end of the text), or somewhere else in the text (cf. pre-contextual versus contextual revisions in Lindgren & Sullivan, 2006).

The features identified from previous studies can be categorized into features related to timings of pauses, revisions, typing bursts, verbosity, and other

Table 1 Descriptive statistics (Mean, SD) and correlational analyses of the (untransformed) keystroke features over the complete writing process [0–30 minutes] and the beginning [0–10], middle [10–20], and end [20–30] of the writing process

			Correlation (Pearson's *r*) final grade			
Keystroke feature	Mean	*SD*	[0–30]	[0–10]	[10–20]	[20–30]
Initial pause time (min)	1.3	1.5	-0.16	-0.16	-0.03	-0.09
Total time (min)	27.4	3.6	-0.01	0.02	0.06	-0.03
Mean IKI	174.7	24.2	-0.10	-0.12	0.01	-0.09
SD IKI	133.9	15.7	0.14	0.24**	0.03	0.18
Median IKI	139.8	22.4	-0.06	-0.04	0.03	-0.05
Largest IKI (min)	0.44	0.20	-0.10	-0.02	0.00	-0.04
Mean IKI within word	146.7	21.6	-0.02	0.04	0.01	-0.04
SD IKI within word	68.4	7.9	0.02	0.03	0.02	-0.01
Mean IKI between words	236.6	54.4	0.03	0.02	0.02	0.04
SD IKI between words	239.8	42.6	-0.11	0.01	-0.06	-0.03
Mean time between words	718.0	205.1	0.02	0.03	0.05	-0.06
SD time between words	894.1	171.7	0.02	-0.03	0.08	-0.01
Mean time between sentences	2954	3279	-0.03	0.03	-0.10	-0.11
SD time between sentences	4195	3850	0.03	0.01	-0.20*	-0.06
Number of IKI 0.5-1 s	133.8	48.9	-0.09	-0.02	-0.11	-0.05
Number of IKI 1–1.5 s	32.5	13.2	-0.09	-0.03	-0.11	-0.04
Number of IKI 1.5-2 s	15.2	7.1	-0.12	-0.14	-0.06	-0.05
Number of IKI 2-3 s	15.7	7.5	-0.12	-0.16	-0.05	-0.02
Number of IKI larger than 3 s	28.5	13.4	0.04	0.01	0.03	0.04
Percentage of long pauses between words	25%	9%	0.22*	0.14	0.20*	0.23*
Number of revisions	89.7	40.5	0.03	0.04	0.03	-0.01
Number of leading-edge revisions	47.4	46.1	0.06	0.05	0.07	0.03
Number of in-text revisions	42.4	29.1	-0.07	-0.02	-0.07	-0.06
Number of backspaces	341.5	172.9	-0.10	-0.07	-0.10	-0.03
Mean time in single backspacing	82.4	18.5	0.15	0.12	0.17	0.12
SD time in single backspacing	21.7	11.9	-0.01	0.04	0.05	0.00
Mean time in multiple backspacing	1598	1046	0.09	0.05	0.09	0.06
SD time in multiple backspacing	3672	3333	-0.02	-0.10	-0.01	-0.16
Percentage of characters final text	54%	18%	-0.11	-0.05	0.03	-0.04
Percentage of characters at leading edge	51%	35%	0.11	0.06	0.07	-0.02
Mean number of keystrokes per burst	15.2	5.2	-0.11	-0.06	-0.14	-0.11
SD number of keystrokes per burst	18.9	8.5	-0.10	-0.03	-0.15	-0.07
Largest number of keystrokes per burst	114.7	72.3	-0.06	-0.02	-0.10	-0.04
Number of bursts	159.1	56.3	0.02	0.08	-0.02	-0.01
Percentage of R-bursts	7%	6%	0.12	0.13	0.10	0.03
Percentage of I-bursts	7%	7%	-0.08	-0.09	-0.07	0.01
Percentage of words in P-bursts	34%	22%	0.04	0.00	0.00	-0.03
Number of production cycles	0.3	0.2	0.15	-0.01	0.15	0.18*
Percentage of linear transitions sentences	73%	8%	-0.13	-0.08	-0.18*	-0.24**

Table 1 (continued)

			Correlation (Pearson's *r*) final grade			
Keystroke feature	Mean	*SD*	[0–30]	[0–10]	[10–20]	[20–30]
Percentage of linear transitions words	11%	12%	0.03	-0.01	0.12	-0.08
Number of keystrokes	2387	811.3	-0.04	0.03	-0.09	-0.02
Number of words	295.1	97.1	-0.05	0.02	-0.09	-0.03
SD number of keystrokes per 30 s	41.1	10.7	-0.03	0.04	-0.10	-0.04
Slope number of keystrokes per 30 s	0.0	0.8	-0.04	-0.03	-0.09	-0.02
Entropy number of keystrokes per 30 s	0.0	0.0	0.03	0.05	0.13	0.05
Uniformity number of keystrokes per 30 s	576.1	168.2	-0.07	0.03	-0.11	-0.04
Local extreme number of keystrokes per 30 s	47.0	7.4	0.12	0.10	0.11	0.04
Mean distance 30 s windows > 1 keystroke	1.3	0.2	-0.07	-0.03	0.01	-0.03
SD distance 30 s windows > 1 keystroke	0.9	0.7	-0.10	-0.04	-0.09	-0.15
Number of focus shifts to translation	0.5	2.7	0.08	0.19*	0.01	0.06
Number of focus shifts to task	19.8	6.6	0.13	0.22*	0.01	0.01
Mean time cut/paste/jump events	545.1	394.7	0.07	0.01	0.12	0.09
SD time cut/paste/jump events	1323	1282	0.11	0.09	0.17	0.18*
Percentage of time spent on other events	50%	16%	0.12	0.03	0.12	0.10

Note. All time-based features are in milliseconds (except stated otherwise); * $p < 0.05$, ** $p < 0.01$. Given the multiplicity, *p*-values should be interpreted as exploratory rather than confirmatory

(non-typing) events. An overview of the features and their descriptive statistics is presented in Table 1 (page 12).

Features related to timings of pauses. The majority of features we extracted are related to the timings of pauses. Most features are related to the IKI: the time from a key press until the next key press. All time-based features showed a large positive skew. Therefore, these features (except for total time, initial time, and maximum IKI) were log transformed, and all values above the 95th percentile were removed. Similar approaches can be found in previous studies (e.g., Grabowski, 2008; Van Waes et al., 2017).

- **Initial pause time**. Time from the start of the keystroke logging (assignment start) until first key press (Allen et al., 2016; Sinharay et al., 2019).
- **Total time**. Time from start of the keystroke logging (assignment start) to last key release (maximum is 30 min; Allen et al., 2016; Deane, 2014; Guo et al., 2018).
- **Mean, Median, *SD*, and maximum IKI**. Metrics of the time from a key press until the next key press (Allen et al., 2016; Sinharay et al., 2019).
- **Mean and *SD* IKI within word**. Metrics of all IKI of keystrokes within words (Deane, 2014; Sinharay et al., 2019).
- **Mean and *SD* IKI between words**. Metrics of all IKI of keystrokes between words, i.e., the last letter of a word and the space between words (Deane, 2014; Sinharay et al., 2019).

- **Mean and *SD* time between words**. Metrics of the time from key press of the last letter of a word until the key press of the first letter of the next word (Deane, 2014; Guo et al., 2018; Zhang et al., 2016).
- **Mean and *SD* time between sentences**. Metrics of the time from key press of the end of a sentence marker until the key press of the first letter of the next sentence (Baaijen & Galbraith, 2018; Deane, 2014).
- **Number of IKI of specific length**. Five features were extracted: the number of IKI between 0.5–1.0 s, 1.0–1.5 s, 1.5–2.0 s, 2.0–3.0 s, and larger than 3.0 s (Allen et al., 2016).
- **Percentage long pauses between words**. Number of pauses between words longer than two *SD* from the mean IKI within the copy task, divided by the total number of pauses between words (Baaijen & Galbraith, 2018).

Features related to revisions. Eight features related to revisions were extracted. Again, all time-based features were log transformed and all values above the 95th percentile were removed.

- **Number of revisions**. Number of insertions away from the leading edge plus the number of sequences of backspaces and delete keystrokes, that do not contain a pause longer than two *SD* from the mean IKI within the copy task, and where the cursor was not moved to a different location in the text during the revision (Barkaoui, 2016).
- **Number of leading-edge revisions**. Number of revisions at the leading edge (pre-contextual revisions; Barkaoui, 2016).
- **Number of in-text revisions**. Number of revisions away from the leading edge (contextual revisions; Barkaoui, 2016).
- **Number of backspaces**. Number of backspaces and delete keystrokes (Allen et al., 2016).
- **Mean and *SD* time in single backspacing**. Metrics of the duration of a sequence of backspaces or delete keystrokes which included only one backspace or delete keystroke (Deane, 2014).
- **Mean and *SD* time in multiple backspacing**. Metrics of the duration of a sequence of backspaces or delete keystrokes which included more than one backspace or delete keystroke (Deane, 2014).
- **Percentage of characters in final text**. The number of characters in the full text, divided by the total number of keystrokes (Baaijen & Galbraith, 2018).
- **Percentage of characters at leading edge**. The number of characters typed at the leading edge of the text, divided by the total number of keystrokes, used as a proxy for the size of contextual revisions (cf. Barkaoui, 2016).

Features related to fluency. Fluency in writing has been argued to be reflected in verbosity (production), process variance, revision, and pausing behavior (Van Waes & Leijten, 2015). Here, we solely focus on the burstiness of the writing. Sentences are argued to be composed in sentence parts, also known as written language bursts (Kaufer et al., 1986). Written language bursts, hereafter ‘bursts’, are defined as sequences of text production without a long pause and without a focus shift (e.g., to the task or an online dictionary). To account for individual differences in typing speed, all bursts are defined as sequences of keystrokes that do not contain pauses longer than two *SD* from the mean IKI within the copy task of the same participant (as in Deane, 2014).

- **Mean, *SD*, and maximum number of characters per burst**. (Sinharay et al., 2019).
- **Number of bursts**. (Sinharay et al., 2019).
- **Percentage of R-bursts**. Number of revision bursts at the leading edge ending in a revision, divided by the total number of bursts (Baaijen et al., 2012).
- **Percentage of I-bursts**. Number of insertion bursts produced away from the leading edge, divided by the total number of bursts (Baaijen et al., 2012).
- **Percentage of words in P-bursts**. Number of words in 'clean' production bursts both initiated and terminated by a long pause (not a revision), divided by the total number of words (Baaijen et al., 2012).
- **Number of production cycles**. Number of groups of bursts without interruptions by *other events* (i.e., all events not resulting in a character being typed, see also features related to other events), divided by the number of words (Baaijen & Galbraith, 2018).
- **Percentage of linear transitions between words**. Number of times the transition to the next word was not interrupted by other events, divided by the total number of transitions between words (Baaijen & Galbraith, 2018).
- **Percentage of linear transitions between sentences**. Number of times the transition to the next sentence was not interrupted by other events, divided by the total number of transitions between sentences (Baaijen & Galbraith, 2018).

Features related to verbosity. Two general features related to verbosity were extracted. In addition, as in Allen et al. (2016), we extracted seven features related to the variability of the keystrokes over time. These features are all related to the number of keystrokes in 30-s time windows. Since the task duration was 30 min, there were a total of 60 time windows.

- **Total number of keystrokes** (Allen et al., 2016; Choi & Deane, 2021).
- **Total number of words** (Likens et al., 2017).
- ***SD* number of keystrokes per 30 s**. Variance of the number of keystrokes in every 30 s window (Allen et al., 2016).
- **Slope of the number of keystrokes per 30 s**. The slope of the linear regression applied to the sequence of keystrokes in every 30 s window (Allen et al., 2016).
- **Entropy of the number of keystrokes per 30 s**. Maximum likelihood estimation of Shannon entropy for the number of keystrokes in every 30 s window, divided by the total number of keystrokes (Allen et al., 2016). Calculated with the 'entropy' function in the R-package 'entropy' (Hausser & Strimmer, 2014).
- **Uniformity of the number of keystrokes per 30 s**. Jensen-Shannon Divergence of a uniform distribution of keystrokes (every window: total number of keystrokes divided by total number of windows) and the actual distribution of

keystrokes per 30 s (Allen et al., 2016). Calculated with the `JSD' function in the R-package 'philentropy' (Drost, 2018).

- **Local extreme number of keystrokes per 30 s**. Number of times the sign of the difference in the number of keystrokes between 30 s window changes, from increasing to decreasing number of keystrokes or vice versa (Allen et al., 2016).
- **Mean and *SD* distance 30 s windows of more than one keystroke**. Distance between 30 s windows with more than one keystroke, giving a measure of the amount and variance of long pauses (Allen et al., 2016).

Features related to other events. Lastly, we also included five keystroke features related to non-typing or '*other'* events to get a broader view of students' writing behavior. Other events are all events which do not result in a character being typed; these include mouse events, insertions, replacements, focus events, control, function, and navigation keys (Baaijen & Galbraith, 2018). Since typos are extremely common, these were not counted as an '*other'* event. We defined a typo as a revision within a word, consisting of a maximum of three backspace or delete keystrokes, and where the IKIs of the delete or backspace keystrokes are shorter than two *SD* from the mean IKI in the copy task.

- **Number of focus shifts to translation or task**. Since focus shifts other than to task or translation (e.g., to social media websites) were rare, we only included those two types of focus shifts: the number of times the focus shifted towards an online dictionary or translation web page, and the number of times the focus shifted towards the task (Leijten et al., 2019).
- **Mean and *SD* cut/paste/jump events**. Time spent on cut (selection followed by a keystroke or insertion), paste (insertion), and jump (mouse click resulting in change in position in the document) events (Deane, 2014).
- **Percentage of time spent on other events**. Time spent on other events, divided by the total time spent (Baaijen & Galbraith, 2018).

To be able to perform early prediction, all features were calculated for the keystrokes up to different timepoints in the writing process. First, the keystroke log was divided into six equal time-based segments. Because the writing task lasted 30 min, this resulted in six segments of five minutes each. Then, all 54 features were calculated for each segment (resulting in six windowed feature sets: 0 – 5, 5–10, … 25–30), as well as up to each segment (resulting in five 'full' feature sets: 0–5, 0–10, … 0–30). For all segments starting after the assignment start (excluding t = 0, e.g., 5–10, 10–15), the segment start time was used instead of the assignment start time to calculate initial pause time and total time. For each model, the windowed- and full-feature sets available up to that timepoint were included. For example, for the model at 15 min, the feature sets from 0–5, 5–10, 10–15, and 0–15 min were included.

The data indicated that four students did not start typing in the first five minutes. For these students, there is no information in the features for the prediction models at five minutes. Therefore, we also determined the performance after a certain number of keystrokes. The keystroke log was divided into six equal segments, each segment containing 1/6th of all keystrokes. Similar to the time-based segments, all 54 features were calculated for each keystroke-based segment and up to each keystroke-based segment.

Analyses

Correlational and machine learning analyses were conducted on the keystroke data. Firstly, correlational analyses were run on all keystroke features to determine which keystroke features identified in previous literature are related to writing quality, measured by final grade (RQ1). These correlational analyses were conducted over both the full writing process, and for three different stages within the writing process (0–10 min, 10–20 min, and 20–30 min).

Thereafter, for each of the time-based and keystroke-based segments discussed in Section 2.4, machine learning models were trained to identify whether keystroke features could be used for timely prediction of writing quality (RQ2). Specifically, keystroke features were trained to predict final grade (regression) and which students might need support (binary classification), see Section 2.7. Given the large number of features (54 features per included segment) and the limited number of observations (126 participants), feature reduction was conducted first (Section 2.6). For the final models, feature importance was calculated to identify which keystroke features are most important for timely writing quality prediction at different stages of the writing process (RQ3).

Feature reduction

This feature reduction process was done for each segment separately, as we expected that the prediction power for the features would differ across the models; that is, certain features might work better at different times in the writing process. The feature reduction process followed the workflow established by Perez-Riverol et al. (2017). First, two filter functions from the R-package ’caret’ (Kuhn, 2019) were applied to the full set of features available at each timepoint. Thus, the resulting feature set was the same for all regression and classification models trained at the same timepoint. In the first filter function, the features with near zero variance were removed (’nearZeroVar’ function). Near zero variance features were defined as the features with a ratio of the most common value to the second most common value smaller than 95/5 or features with less than 10% unique values. In the second filter function, highly correlated features with pair-wise correlations above 0.80 were identified (’findCorrelation’ function). From these pairs of highly correlated features, the feature with the largest mean absolute correlation with the other features was removed.

Based on the filtered feature set, a wrapper function from the R-package ’caret’ (Kuhn, 2019) was conducted using tenfold cross-validation. A wrapper function reduces the features based on their effect on the model performance. Specifically, recursive feature elimination was applied as the wrapper function (’rfeControl’ function). With recursive feature elimination, all features are first used to fit the model, and the features are ranked according to their importance. At every following step, the model is fitted again with all features except for the predictor that had the lowest importance (according to the previous step). This process is repeated recursively. Here, the best subset of features was determined using tenfold cross-validation by selecting the subset of features that resulted in the lowest root mean squared error (RMSE) for the regression models, and the largest AUC for the classification models. In addition, to avoid overfitting, a simpler model (fewer features) was preferred over a more complex model if the difference in RMSE was less than 1% ('pickSizeTolerance' function). Thus, the final sets of features used in the prediction differed for each model. However, as the feature reduction approach was identical for all models, comparing models for a specific timepoint and across timepoints is still valid.

Machine learning algorithms

Regression models were trained on the keystroke features to predict final grade at different stages in the writing process. Six timepoints were based on the elapsed time (every five minutes within the writing process), and six other timepoints were based on the number of keystrokes (every 1/6$^{\text{th}}$ of the total number of keystrokes). Three regression models were run with the data available at each of those timepoints: random forest, support vector machines with radial kernel, and naive Bayes. These models are suitable for the small number of observations in our data. Random forests and support vector machines were chosen as they generally work well on continuous data. In addition, naive Bayes was added to determine if a simpler model would work equally well. The mean final grade was used as a baseline (Mean grade = 5.5, see Section 2.3). Root mean squared error (RMSE) was used as an evaluation metric. We favored RMSE over mean absolute error (MAE), as we wanted to assign a larger penalty to larger errors. Parameter tuning was using tenfold cross-validation via the built-in parameter tuning function within the 'trainControl' function, optimizing for the smallest RMSE. Within each of these ten folds, feature reduction was conducted as described in Section 2.6.

In addition, binary classification models were trained on the keystroke features to predict students who might need support at different points in the writing process. Participants with a score lower than 5.5 (on a scale from 1–10) were classified as 'might need support', and those with a score equal to or higher than 5.5 were classified as 'do not need support'. Three classification models—random forest, support vector machines with radial kernel, and naive Bayes—were run with the data available at the same timepoints as used in the regression models (six timepoints based on elapsed time and six timepoints based on the number of keystrokes). The majority class (52% would not directly need support, see Section 2.3) was used as a baseline. AUC, precision, recall, and F_1-score were used as evaluation metrics. Parameter tuning was done using tenfold cross-validation via the built-in parameter tuning function within the 'trainControl' function, optimizing for the largest F_1-score. Similarly to the regression models, feature reduction was conducted within each of the ten folds, as described in Section 2.6.

Results

Relationship between keystroke data and writing quality

To determine which keystroke features identified in previous literature are related to writing quality (RQ1), correlational analyses were conducted. Table 1 presents the correlations of all keystroke features with final grade. For the keystrokes measured over the full writing processes, only one significant correlation was found. Final grade had a small positive relation to the percentage of long pauses between words ($r = 0.22$). Thus, more pauses between words are related to a higher writing quality. The correlations over the beginning [0–10 min], middle [10–20 min], and end [20–30 min] of the writing process showed different patterns. The *SD* of the IKI ($r = 0.24$) and the number of focus shifts to translation ($r = 0.19$) and task ($r = 0.22$) were significantly related to final grade only in the beginning of the writing process. In contrast, the *SD* of the time between sentences ($r = -0.20$), percentage of long pauses between words ($r = 0.20$ and $r = 0.23$), number of production cycles ($r = 0.18$), percentage of linear transitions between sentences ($r = -0.18$ and $r = -0.24$), and the *SD* of the time in cut, paste, and jump events ($r = 0.18$) were only significantly related to final grade in the middle or end of the writing process.

To conclude, this indicates that the relationship between keystroke data and writing quality is limited. In addition, this shows that the relation of the features with final grade differs over time within the writing process.

Timely writing quality prediction

To identify whether keystroke features can be used for timely writing quality prediction (RQ2), we first predicted final grade (regression) at the different timepoints (based on time elapsed and number of keystrokes) in the writing process. None of the models outperformed the baseline (mean final grade) at any timepoint in the writing process. Thus, the keystroke features cannot be used for the early prediction of final grade, nor for the prediction of final grade once the writing process is finished.

Yet, for identifying students who might need support, we would not need to predict such a specific outcome variable. Rather than predicting final grade, it would be enough to predict whether or not the student would pass the assignment. Therefore, we also predicted which students might need support (final grade < 5.5; binary classification). Figure 1 shows the performance of the classification models predicting which students might need support after every five minutes of keystroke data. First, it shows that the models only occasionally outperform the majority class baseline. We expected that the error would decrease as more information becomes available (when more minutes of keystroke data are used in the model). However, this trend was not clearly visible. Lastly, there does not seem to be one classification model that outperforms the other models at all points in time during the writing process.

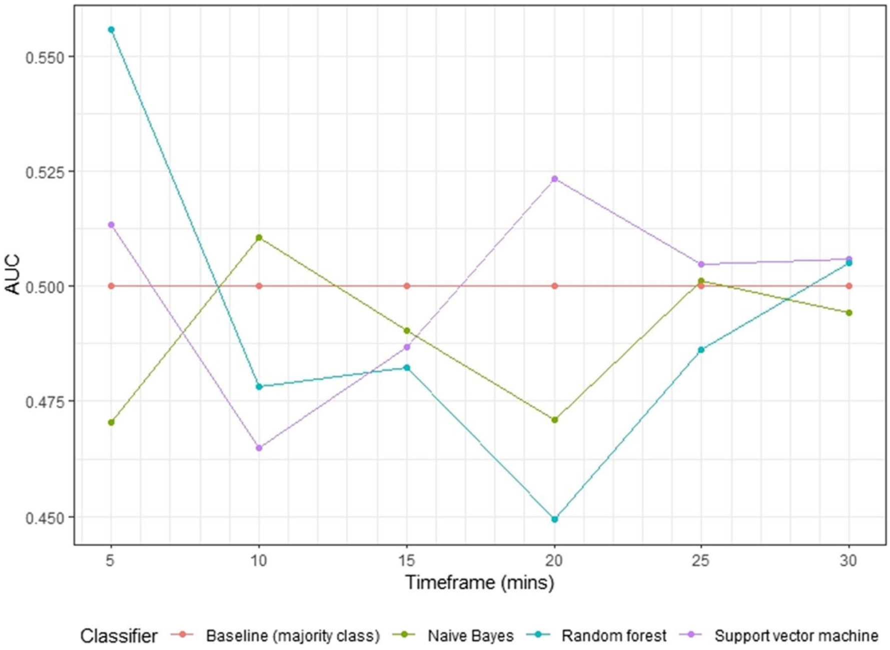

Fig. 1 AUC of the classification models predicting which students might need support, compared to the baseline, for the keystroke data up to every five minutes

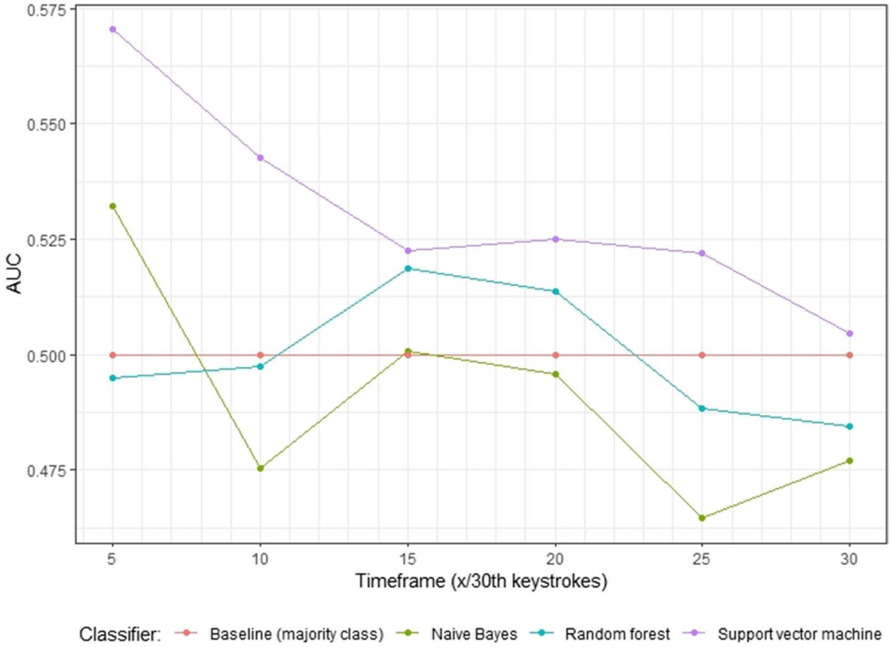

Fig. 2 AUC of the classification models predicting students who might need support, compared to the baseline, for the keystroke data up to every x/30th of the total number of keystrokes

Figure 2 shows the performance of the classification models predicting which students might need support after every 1/6th of the total number of keystrokes. These models appear to perform slightly better than the models for every five minutes. The support vector machine outperforms the baseline and all other models at every timepoint within the writing process. Hence, the support vector machine appears to be the best model tested. However, with an AUC of 0.57, these models still do not perform *well*. Interestingly, the models also do not seem to improve over time. For the naive Bayes and random forest, there seems to be an increase at 15/30$^{\text{th}}$ and 20/30$^{\text{th}}$ of the total amount of keystrokes, but when more keystrokes are added, the performance decreases again. Lastly, for the support vector machine and naive Bayes classifier, the AUC is highest when only 5/30$^{\text{th}}$ of the total amount of keystrokes are included in the model. This would indicate that the other keystrokes do not add much additional information for the prediction of which students might need support. To conclude, this showed that the keystroke features obtained from previous literature have limited value for (early) writing quality prediction (regression and classification) in our current sample.

Feature importance for timely writing quality prediction

To get insight into which features had the highest predictive value and whether the most important features differed at the different timepoints (RQ3), we calculated the feature importance for all classification models. The feature importance for each model (up to the five best features) for every five minutes is shown in Fig. 3. Here, the columns show the three prediction models including data up to the different time points (e.g., 0–05 shows the prediction after the first 5 min of the writing process). The rows show the top five most important features in these models, including the timeframe over which each feature was calculated (displayed between brackets).

For the first five minutes of keystroke data, the number of focus shifts to task, the distance between 30 s windows with more than one keystroke (variance in long pauses), and the initial pause time were the most important features. When more data were considered (from 10 up to 30 min), these features were less important. Only the number of focus shifts to task in the first five minutes were still somewhat important in the later models. In the middle of the writing process, the percentage of other events (5–10 min), the *SD* of the IKI within words (10–15 min), and the number of production cycles (5–10 min) were most important. At the end of the writing process (20–30 min), no features seemed to be of specific importance. The *SD* of the IKI between words appeared to be most important when measured over the largest possible timeframe (e.g., 0–25 and 0–30 min).

The feature importance for each model for every x/30$^{\text{th}}$ of the total number of keystrokes is shown in Fig. 4. For most of the models, the mean and *SD* of the IKI of the first 1/6$^{\text{th}}$ of the total number of keystrokes were most important. In addition, features related to revisions in the beginning and the middle of keystroke production were found to be important, such as the percentage of revision bursts (2–3/6$^{\text{th}}$ of the keystrokes), mean time in multiple backspacing (0–1/6$^{\text{th}}$ of the keystrokes), and mean time in cut/paste/jump events (1–2/6$^{\text{th}}$ of keystrokes).

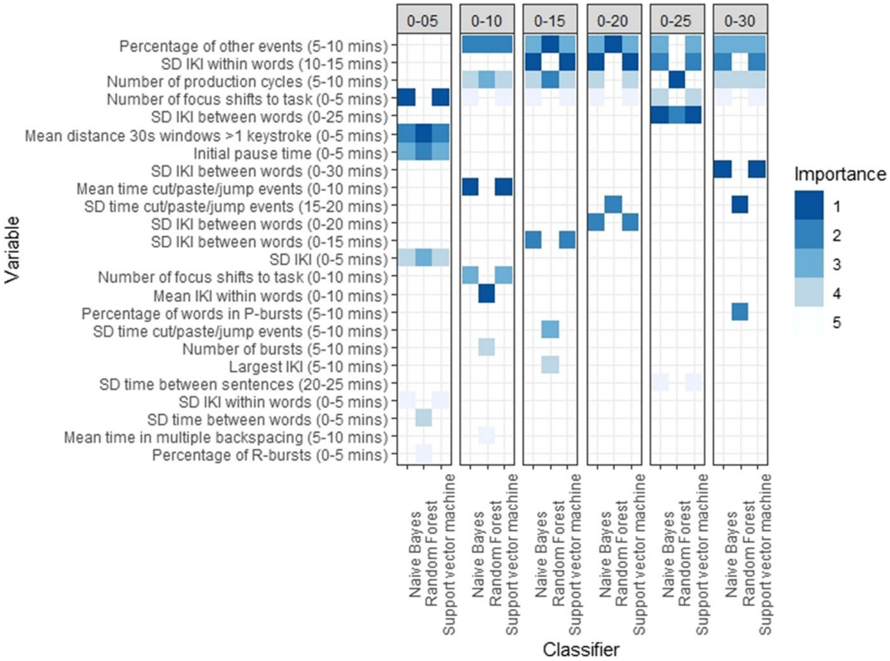

Fig. 3 Feature importance for each model predicting who might need support, using the keystroke data up to every five minutes. *Note*. Features are ordered by importance; 1 is the best feature, 2 is the second-best feature, etc., across all resamples in the recursive feature elimination. Only the five best features per model (or fewer if the best model consisted of fewer features) are listed

No features seemed to be consistently important in the last segments of keystrokes (3–6/6th of keystrokes). In addition, no features measured over the full writing process seemed to be of particular importance.

As the models showed rather low predictive power, the feature importance needs to be interpreted with caution. However, the feature importance was relatively consistent over the different models, which indicates there are still some features that show an effect irrespective of the classifier. Thus, there might still be some predictive power within the features. That said, the most important features differed between the models using the keystroke data up to every five minutes versus up to every x/30th keystroke. In addition, the figures show that the feature importance varied per timeframe over which the feature was calculated. For example, the percentage of ‘other’ events were only important when measured in the middle of writing process (5–10 min), while the number of focus shifts was only important when measured in the beginning (0–5 min). Lastly, the feature importance was shown to differ somewhat over time. For example, for early prediction (after 5 or 10 min), the *SD* IKI between words was not important, while it was one of the most important predictors for prediction late in the writing process (after 25 or 30 min).

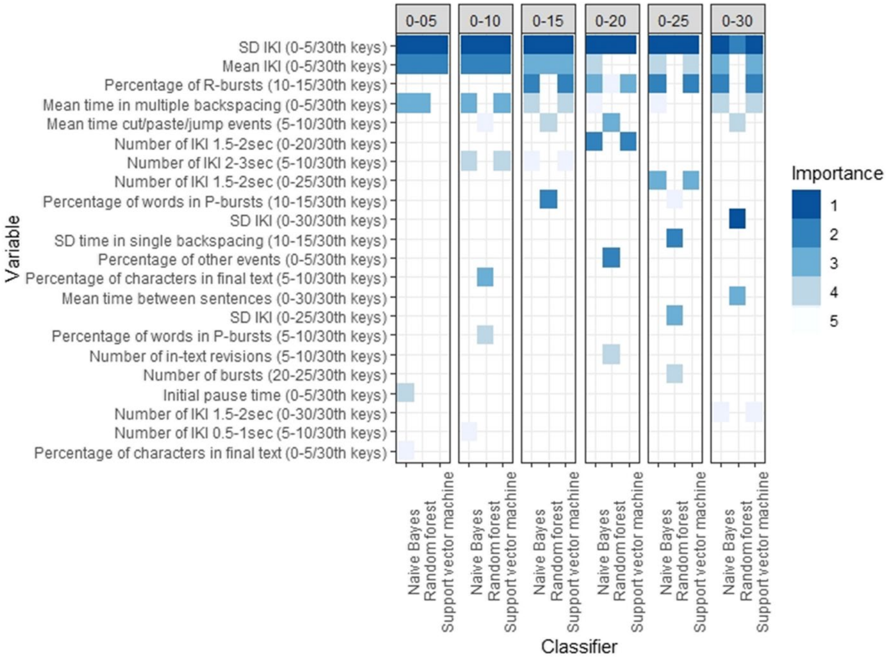

Fig. 4 Feature importance for each model predicting who might need support, using the keystroke data up to every x/30$^{\text{th}}$ of the total number of keystrokes. *Note*. Features are ordered by importance; 1 is the best feature, 2 is the second-best feature, etc., across all resamples in the recursive feature elimination. Only the five best features per model (or fewer if the best model consisted of fewer features) are listed

Discussion

In this study, we aimed to further explore the relationship between keystroke data and writing quality. Specifically, we wanted to identify (RQ1) which keystroke features identified in previous research are related to writing quality, (RQ2) whether these can be used for timely writing quality prediction at different stages in the writing process, and (RQ3) which features are important for these prediction models.

Relation between keystroke data and writing quality

Based on previous research on writing quality prediction using keystroke data, 54 keystroke features were identified. First, correlational analyses were conducted to identify the relationship between these features and writing quality. The correlation analyses showed that only the percentage of long pauses between words, measured over the full writing process, was significantly related to final grade: more pauses between words resulted in higher grades. This is in contrast with previous studies, which showed that each of the 54 keystroke features at some point have been related to final grade. For example, Allen et al. (2016) found medium to high effect sizes for the correlations between essay scores and the number of words, median IKI, entropy, and extremes of the number of keystrokes per 30 s. Sinharay et al. (2019) also found medium correlations between time within words and number of bursts. None of these were found here.

The differences in correlations can be explained by the different tasks used in these studies. Guo et al. (2018) showed that these correlations already differed for similar tasks with different prompts. Here, the differences were even larger; this might be because the current study analyzed a summary writing task, while previous studies analyzed argumentative and persuasive essay writing tasks (see e.g., Sinharay et al., 2019; Zhang et al., 2016). The different tasks might lead to different writing processes, and consequently also different keystroke data. Previous work already showed that keystroke data is sensitive to small differences in writing tasks, such as copy writing versus email writing (Conijn et al., 2019). In addition, differences in source usage, such as independent and integrated tasks, have been shown to have an effect on students' pausing and revision behavior (Michel et al., 2020). Therefore, the relationship between keystroke data and writing quality might also differ across tasks.

This can be further exemplified by some specific keystroke features. For example, previous work showed that especially features related to verbosity, such as total time, number of words, and number of keystrokes, resulted in moderate to high positive correlations with writing quality (e.g., Allen et al., 2016; Sinharay et al., 2019). Choi and Deane (2021) even found that the number of keystroke features was the only consistent predictor across four writing tasks. Features related to verbosity were not found to be significantly related in the current study. This can be explained by differences in the task requirements. Contrary to previous studies, the present task requested students to write within a specified word limit, which a majority of the students did (only 21% of the students wrote more than 10% over the word limit). Accordingly, writing more did not usually result in higher grades; it showed that the student did not comply with the requirements, and often resulted in irrelevant information added, which usually resulted in lower grades. Additionally, some previous studies used a non-timed essay task (e.g., Allen et al., 2016; Guo et al., 2018), while in our study, participants were asked to write the academic summary in 30 min. Therefore, there is presumably less variance in some of the features, such as total time, number of words, and number of revisions, resulting in a lower predictive power of these features. This result suggests that relative features need to be used (e.g., number of revisions per word) to avoid task length effects. Moreover, the fact that verbosity or simply 'writing more' results in higher writing quality as found in previous work, has limited value for pedagogical applications.

Timely writing quality prediction

After the correlational analysis, regression analyses were conducted to model the relationship between the keystroke features and final grade. Not surprisingly, the low correlations also resulted in low prediction accuracies. Although previous studies were able to predict final grade to a large extent (e.g., Allen et al., 2016; Sinharay et al., 2019), none of our models were able to predict final grade with a higher performance than the baseline. These divergent findings can again be explained by the differences in tasks between the current study and previous studies. In addition, this shows that models of writing quality prediction might not generalize across tasks. However, for the timely identification of students who might need support, we do not need to predict the exact final grade. Therefore, binary classification was used to determine whether the student might need support at different timepoints in the writing process.

For the early prediction of which students might need support, keystroke data were included up to different timepoints within the writing process. Two approaches were used to identify these different timepoints based on time elapsed and the number of keystrokes. However, the accuracy was low in all cases: the models only slightly outperformed the baseline at some of the timepoints. Thus, these features do not allow for the early identification of students who might need support. In addition, although the performance was expected to improve over time as more data was added, this was not the case. This could probably be explained by the fact that the correlation with final grade at the end of the writing process was already limited, providing little room for the performance to increase over time.

The models using the timepoints based on the number of keystrokes resulted in the best performance, with the support vector machine model consistently outperforming the baseline. This might be because by dividing the process according to the number of keystrokes, there is always information within these segments. In contrast, when dividing the process by the time elapsed, there might be limited information in the first and last segment, because the writer did not yet start, or already finished their writing, respectively. Thus, although many approaches currently divide the writing process based on time (e.g., Zhang et al., 2016), the amount of information available in these segments should be taken into account as well.

Feature importance for timely writing quality prediction

Lastly, we wanted to determine which features were most important, and how this feature importance differed over time. Given that the performance of the models was rather low, the results need to be interpreted with caution. However, the mean and *SD* IKI of the first 5/30 = 16.7% of the keystrokes and the number of production cycles and percentage of other events in the 5th to 10th minute show some robust effects, which might indicate that these features still contain some predictive power.

In addition, the feature importance analysis showed that certain features were only relevant in the middle of the writing process (e.g., percentage of other events or percentage of revision bursts), while others are only important in the beginning (e.g., number of focus shifts to task or mean and *SD* IKI). These findings corroborate previous work, which also found that the relationship between keystroke features and quality differs over time, in both size and the direction of the effect (Breetvelt et al., 1994). Future work should further examine *how* the keystroke features change over time, for example by using pattern mining techniques or temporal analysis. Moreover, this might indicate that a single keystroke feature might be rather sensitive to fluctuations in the writing process, further stressing the importance of examining keystroke features grouped into factors (Galbraith & Baaijen, 2019).

Limitations

This study is limited in some ways. First, we included keystroke features which were previously found important for the prediction of writing quality once the writing process has finished. However, we showed that the importance of these features for writing quality prediction differed over time in the writing process. Therefore, these features might not be the best features for the *early* prediction of writing quality. Future work should concentrate on identifying specific features that would be more informative for early prediction. In addition, to further improve prediction accuracy, future work should also include information on writing profiles to account for individual differences in writing preferences (Levy & Ransdell, 1996; Van Waes & Schellens, 2003). Moreover, future work could focus on creating higher-level features – for example, by combining the keystroke data with information derived from natural language processing – to aggregate at the word, phrase, or sentence level (Leijten & Van Waes, 2013).

Secondly, we included keystroke features which were previously found to be important for the prediction of writing quality, but demonstrated in tasks highly different from the current task. It could therefore be that these keystroke features do not generalize well to the current task. The literature is also unclear on how the relationship between keystroke data and writing quality differs across tasks. For example, Deane (2014) showed that process and product features of one task could be used to predict another writing task, which indicates that some generalization between tasks is possible. However, Guo et al. (2018) showed that the variance explained by keystroke features differed across six tasks. Therefore, future work should further investigate how generalizable the used keystroke features are for the prediction of writing quality across tasks and, specifically of interest for the current study, how generalizable the features are in the context of early prediction of writing quality across tasks.

Lastly, the current approach aims to identify students who might need support. However, this is only the first step in helping these students (Romero & Ventura, 2019; Sonderlund et al., 2019). The current approach does not provide any information on *why* a student needs support. Hence, this provides limited insight into the best timing, content, or design of the feedback or computer-based writing support. Although this is out of the scope of the current study, future experimental studies could determine the effect of the content and the design of the feedback on the early writing quality predictions. The keystroke data might also be used to identify why a student needs support while also generating the content of the writing feedback. Implications for this based on the current findings are discussed below.

Implications for automated writing process analysis and future research

Current computer-based writing support systems mostly assess and provide feedback on the writing product, as opposed to the writing process (Cotos, 2015; Ma et al., 2014; Wang et al., 2013). Here, *timely* feedback is especially important for feedback to be effective (Cotos, 2015; Ferguson, 2011; Shute, 2008). In this article, we determined whether keystroke data could be used to predict writing quality as soon as possible, in order to identify the students who would need support. Even though we included features that were shown to be related to writing quality in previous studies, our model is not accurate enough for the early prediction of writing quality. Thus, this might not be directly used to provide support to the students. However, we contend that automated analysis of the writing process may still be useful for writing development. Therefore, based on the current study, we briefly explore three themes for future research on automated analysis of keystroke data that could further inform us about students' writing processes and the effects on writing development.

First, we feel it is important to note that there is no such a thing as a 'single' writing process that results in high quality for all students and writing tasks at all times. In the current study, we showed that models for predicting writing quality might not be generalizable across tasks. Moreover, writers have individual preferences for specific processes or approaches, which can also be influenced by the task (Levy & Ransdell, 1996; Van Waes & Schellens, 2003). These preferences are also called writing profiles, or writing signatures. Different writing profiles might benefit from different types of feedback at different stages in the writing process. Clustering techniques may be used in future work to automatically extract these writing profiles and identify how these profiles relate to writing quality (Uto et al., 2020).

Second, the information on the writing process as obtained by keystroke features can be used to visualize these writing profiles in so called process or progression graphs (Leijten & Van Waes, 2013; Perrin, 2019), which can be subsequently be used for feedback purposes. It has been quite difficult to create a pedagogical context in which students were challenged to reflect on their writing process. However, given the current developments in keystroke analysis, we can automatically present students with both detailed and wide perspectives on different aspects of their writing process, e.g., based on the features extracted in the current study (related to their pausing or revision behavior, fluency, or source usage—for an example of a process feedback report, see Vandermeulen et al., 2020). Although these might not provide 100% accurate predictions of students' writing quality, they might still be used by students to reflect on their process. By challenging students to reflect on their writing process and (possibly) compare their results with their peers, we argue that we will be able to create a more solid basis to feed students' process awareness, encourage self-assessment and foster self-regulation.

The insights into students' writing processes using keystroke data can be visualized in so-called learning dashboards, or writing dashboards specifically, that provide an overview of students' writing activities over time (Conijn et al., 2020a, b; Verbert et al., 2014). For example, process graphs have been used to visualize students' text production over time (Leijten & Van Waes, 2013). Figure 5 provides an example of two writers from the current dataset with similar writing quality. Here, one writer plans extensively at the beginning of the writing process, with many copy-paste actions (left). In contrast, the other writer skips planning in the beginning, but revises extensively at a later stage in the writing process, indicated by the

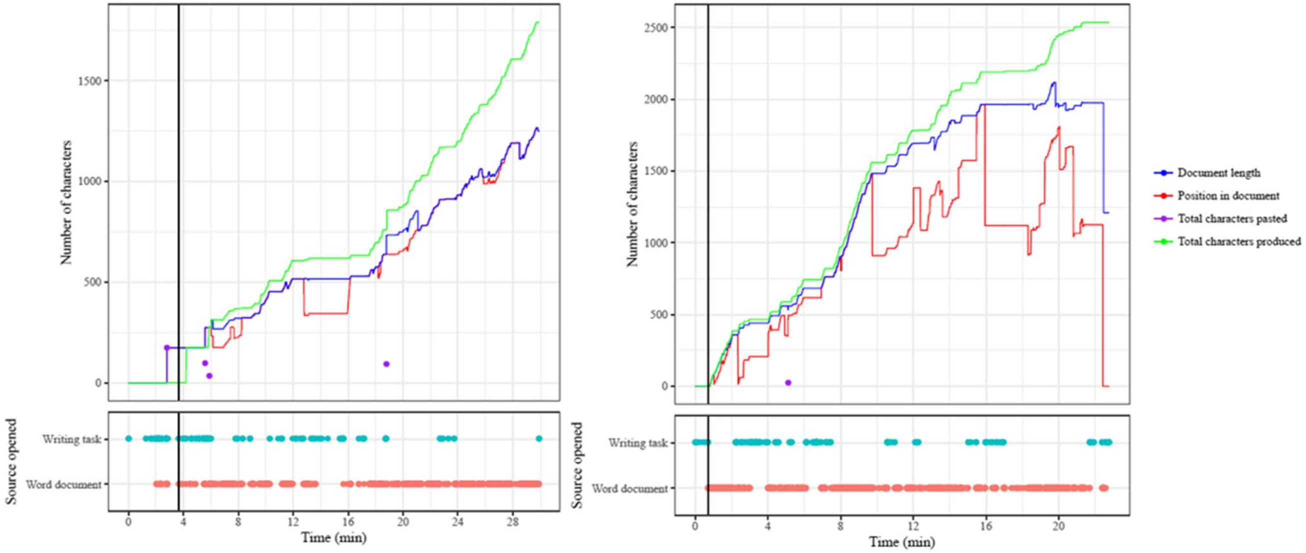

Fig. 5 Progress graph of two students with similar grades, showing different writing profiles. Left: extensive planning, little revision (participant 10, final grade 9.5); Right: little planning, extensive revision (participant 59, final grade 9)

changing position in the document (see red line; right). Although this results in different writing process characteristics, these writers still show similar writing quality. This illustrates that there might not be a single ’best’ writing process. A small case study showed that a feedback report including the process graph helped students to reflect on their writing process and even enabled them to identify ways to improve their writing process (Vandermeulen et al., 2020).

For these writing dashboards to be effective, it is important to ensure that they result in *actionable* feedback (cf. Conde & Hernández-García, 2015). Recently, human-centered approaches have been suggested as a solution to more actionable and effective feedback systems (Buckingham Shum et al., 2019), focusing on iterative design with teachers and students (Conijn et al., 2020a, b; see e.g., Martinez-Maldonado et al., 2015; Wise & Jung, 2019). Recently, it was shown that these approaches indeed could be used to determine potential actions upon writing dashboards, such as reflection on students’ writing processes (Conijn et al., 2020a, b). Future work should further identify how writing process data may be visualized to provide actionable feedback and further foster students’ reflection on the writing process to improve students’ writing.

Conclusion

This study provided insight into the relationship between the writing process, measured by keystroke features, and writing quality, measured by final grade. In addition, we determined which of these features are useful for predicting writing quality *during* the writing process. In contrast to previous research, the relationship between the keystroke features and writing quality proved to be rather limited. Therefore, it was not possible to predict writing quality, let alone early prediction. Hence, the features identified in previous work for predicting writing quality might not generalize across writing tasks and writing contexts. In addition, we showed that the relationship between keystroke features and writing quality changes over time during the writing process. To conclude, this paper shows that the relationship between keystroke features and writing quality might be less straight-forward than originally posited. We contend that more in-depth analyses of keystroke data may still be used to inform us about students’ writing processes and improve students’ writing.

Appendix A Rubric academic summary

	1	2	3	4
1. Main idea	Main idea is not present	Main idea is not clearly stated	Main idea is mostly clear; or main idea is clear, but not within the first two sentences	Main idea is clear and within the first two sentences

	1	2	3	4
2. Structure and organization	The paragraph is poorly structured and hard to follow	The paragraph lacks some structure, and could be more cohesive and/or consistent	The paragraph is well-structured, but could be more cohesive and/or consistent	The paragraph is well-structured, cohesive, and consistent
3. Content	More than one piece of critical supporting evidence is missing; or instances of incorrect material	Critical supporting evidence is missing; or many instances of irrelevant material	Most supporting evidence is included; some irrelevant material	All supporting evidence is included; no incorrect or irrelevant material
4. Language and paraphrasing	Author plagiarizes	Author uses quotes or sentences too similar to the text	Author uses own words, but limited variety in words and sentence structures	Author uses own words, wide variety in words and sentence structures
5. Grammar and mechanics	> 5 spelling and/or grammar mistakes	3–4 spelling and/or grammar mistakes	1–2 spelling and/or grammar mistakes	No spelling and/or grammar mistakes

Task compliance (points subtracted):

No in-text citation: 0.5

Between 200 and 300 words: 0.5

Less than 100 or more than 300 words: 1.0

Final grade = (total score rubrics / 2) - task compliance.

References

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Computer-based writing instruction. *Handbook of Writing Research*, 316–329.

Allen, L. K., Jacovina, M. E., Dascalu, M., Roscoe, R. D., Kent, K., Likens, A. D., & McNamara, D. S. (2016). {ENTER}ing the Time Series {SPACE}: Uncovering the Writing Process through Keystroke Analyses. *Proceedings of the 9th International Conference on Educational Data Mining (EDM)*, 22–29. https://eric.ed.gov/?id=ED592674.

Baaijen, V. M., & Galbraith, D. (2018). Discovery Through Writing: Relationships with Writing Processes and Text Quality. *Cognition and Instruction, 36*(3), 1–25. https://doi.org/10.1080/07370008.2018.1456431.

Baaijen, V. M., Galbraith, D., & de Glopper, K. (2012). Keystroke Analysis: Reflections on Procedures and Measures. *Written Communication, 29*(3), 246–277. https://doi.org/10.1177/0741088312451108.

Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal, 100*(1), 320–340. https://doi.org/10.1111/modl.12316.

Beauvais, C., Olive, T., & Passerault, J.-M. (2011). Why are some texts good and others not? Relationship between text quality and management of the writing processes. *Journal of Educational Psychology, 103*(2), 415. https://doi.org/10.1037/a0022545.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*(3), 191–205. https://doi.org/10.1016/j.jslw.2005.08.001.

Bixler, R., & D'Mello, S. (2013). Detecting Boredom and Engagement During Writing with Keystroke Analysis, Task Appraisals, and Stable Traits. *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, 225–234. https://doi.org/10.1145/2449396.2449426.

Braaksma, M. A., Rijlaarsdam, G., Van den Bergh, H., & van Hout-Wolters, B. H. M. (2004). Observational learning and its effects on the orchestration of writing processes. *Cognition and Instruction, 22*(1), 1–36. https://doi.org/10.1207/s1532690Xci2201_1.

Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction*, *12*(2), 103–123. JSTOR. https://www.jstor.org/stable/3233677.

Buckingham Shumhandler, S., Ferguson, R., & Martinez-Maldonado, R. (2019). Human-centred learning analytics. *Journal of Learning Analytics, 6*(2), 1–9. https://doi.org/10.18608/jla.2019.62.1.

Casey, K. (2017). Using keystroke analytics to improve pass-fail classifiers. *Journal of Learning Analytics, 4*(2), 189–211. https://doi.org/10.18608/jla.2017.42.14.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*(3), 267–296. https://doi.org/10.1016/S1060-3743(03)00038-9.

Choi, I., & Deane, P. (2021). Evaluating Writing Process Features in an Adult EFL Writing Assessment Context: A Keystroke Logging Study. *Language Assessment Quarterly, 18*(2), 107–132. https://doi.org/10.1080/15434303.2020.1804913.

Choi, Y. H. (2007). On-line revision behaviors in EFL writing process. *English Teaching, 62*(4), 69–93.

Conde, M. Á., & Hernández-García, Á. (2015). Learning analytics for educational decision making. *Computers in Human Behavior, 47*, 1–3. https://doi.org/10.1016/j.chb.2014.12.034.

Conijn, R., Roeser, J., & van Zaanen, M. (2019). Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing, 32*(9), 2353–2374. https://doi.org/10.1007/s11145-019-09953-8.

Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van Waes, L., & Van Zaanen, M. (2020a). How to provide automated feedback on the writing process? A participatory approach to design writing analytics tools. *Computer Assisted Language Learning*, 1–31. https://doi.org/10.1080/09588221.2020.1839503.

Conijn, R., Van Waes, L., & van Zaanen, M. (2020b). Human-centered design of a dashboard on students' revisions during writing. *Conference Proceedings of the 14th European Conference on Technology Enhanced Learning, EC*-TEL, 1–15. https://doi.org/10.1007/978-3-030-57717-9_3.

Cotos, E. (2015). Automated Writing Analysis for writing pedagogy: From healthy tension to tangible prospects. *Writing and Pedagogy, 6*, 1. https://doi.org/10.1558/wap.v7i2-3.26381.

Crossley, S. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3). https://doi.org/10.17239/jowr-2020.11.03.01.

Deane, P. (2014). Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *ETS Research Report Series, 2014*(1), 1–23. https://doi.org/10.1002/ets2.12002.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1). http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640.

Drost, H.-G. (2018). *philentropy: Similarity and Distance Quantification Between Probability Functions (R package version 0.3.0)*. https://CRAN.R-project.org/package=philentropy.

Emerson, A., Smith, A., Smith, C., Rodríguez, F. J., Min, W., Wiebe, E. N., Mott, B. W., Boyer, K. E., & Lester, J. C. (2019). Predicting Early and Often: Predictive Student Modeling for Block-Based Programming Environments. *Proceedings of the 12th International Conference on Educational Data Mining*, 39–48.

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*(4), 400–414.

Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education, 36*(1), 51–62. https://doi.org/10.1080/02602930903197883.

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication, 31*(1), 21–32.

Galbraith, D., & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In E. Lindgren & K. Sullivan (Eds.), *Observing Writing* (Vol. 38, pp. 306–325). Brill. https://doi.org/10.1163/9789004392526_015.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007.

Grabowski, J. (2008). The internal structure of university students' keyboard skills. *Journal of Writing Research*, *1*(1). https://doi.org/10.17239/jowr-2008.01.01.2.

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.

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.

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

Hausser, J., & Strimmer, K. (2014). *entropy: Estimation of Entropy, Mutual Information and Related Quantities (R package version 1.2.1)*. https://CRAN.R-project.org/package=entropy.

Kaufer, D. S., Hayes, J. R., & Flower, L. (1986). Composing written sentences. *Research in the Teaching of English*, 121–140.

Kellogg, R. T. (1987). Writing performance: Effects of cognitive strategies. *Writing performance, 4*(3), 269–298. https://doi.org/10.1177/0741088387004003003.

Knight, S., Shibani, A., Abel, S., Gibson, A., & Ryan, P. (2020). AcaWriter: A Learning Analytics Tool for Formative Feedback on Academic Writing. *Journal of Writing Research, 12*(1), 141–186. https://doi.org/10.17239/jowr-2020.12.01.06.

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

Kuhn, M. (2019). *caret: Classification and Regression Training (R package version 6.0–84)*. https://CRAN.R-project.org/package=caret.

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.

Leijten, M., Van Waes, L., Schrijver, I., Bernolet, S., & Vangehuchten, L. (2019). Mapping master's students' use of external sources in source-based writing in L1 and L2. *Studies in Second Language Acquisition, 41*(3), 555–582. https://doi.org/10.1017/S0272263119000251.

Levy, C. M., & Ransdell, S. (1996). Writing signatures. In C. M. Levy & S. Ransdell, *The science of writing: Theories, methods, individual differences and applications*. Lawrence Erlbaum Associates, Inc.

Likens, A. D., Allen, L. K., & McNamara, D. S. (2017). Keystroke Dynamics Predict Essay Quality. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (CogSci 2017)*, 2573–2578.

Lindgren, E., & Sullivan, K. P. (2006). Analysing online revision. In *Computer Keystroke Logging and Writing: Methods and Applications* (pp. 157–188). Elsevier. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A143735.

Lindgren, E., & Sullivan, K. P. (2019). *Observing Writing: Insights from Keystroke Logging and Handwriting*. Brill. https://doi.org/10.1163/9789004392526.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*(4), 901–918. https://doi.org/10.1037/a0037123.

Macfadyen, L., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*(2), 588–599. https://doi.org/10.1016/j.compedu.2009.09.008.

Mao, Y., Zhi, R., Khoshnevisan, F., Price, T., Barnes, T., & Chi, M. (2019). One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. *Proceedings of the 12th International Conference on Educational Data Mining*, 119–128.

Martinez-Maldonado, R., Pardo, A., Mirriahi, N., Yacef, K., Kay, J., & Clayphan, A. (2015). LATUX: An iterative workflow for designing, validating and deploying learning analytics visualisations. *Journal of Learning Analytics, 2*(3), 9–39. https://doi.org/10.18608/jla.2015.23.3.

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.

Michel, M., Révész, A., Lu, X., Kourtali, N.-E., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research*, 0267658320915501.

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*(2), 68–85. https://doi.org/10.1016/j.asw.2010.05.004.

Passonneau, R. J., McNamara, D., Muresan, S., & Perin, D. (2017). Preface: Special issue on multidisciplinary approaches to AI and education for reading and writing. *International Journal of Artificial Intelligence in Education, 27*(4), 665–670. https://doi.org/10.1007/s40593-017-0158-8.

Perez-Riverol, Y., Kuhn, M., Vizcaíno, J. A., Hitz, M.-P., & Audain, E. (2017). Accurate and fast feature selection workflow for high-dimensional omics data. *PLoS ONE, 12*(12), e0189875.

Perrin, D. (2019). Progression analysis: Working with large data corpora in field research on writing. In *Observing Writing: Insights from Keystroke Logging and Handwriting* (pp. 143–162). Brill. https://doi.org/10.1163/9789004392526_008.

Polyzou, A., & Karypis, G. (2019). Feature extraction for next-term prediction of poor student performance. *IEEE Transactions on Learning Technologies, 12*(2), 237–248. https://doi.org/10.1109/TLT.2019.2913358.

Rapp, C., & Kauf, P. (2018). Scaling Academic Writing Instruction: Evaluation of a Scaffolding Tool (Thesis Writer). *International Journal of Artificial Intelligence in Education, 28*(4), 590–615. https://doi.org/10.1007/s40593-017-0162-z.

Romero, C., & Ventura, S. (2019). Guest editorial: Special issue on early prediction and supporting of learning performance. *IEEE Transactions on Learning Technologies, 12*(2), 145–147. https://doi.org/10.1109/TLT.2019.2908106.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795.

Sinharay, S., Zhang, M., & Deane, P. (2019). Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education, 32*(2), 116–137. https://doi.org/10.1080/08957347.2019.1577245.

Sonderlund, A., Hughes, E., & Smith, J. R. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology, 50*(5), 2594–2618. https://doi.org/10.1111/bjet.12720.

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

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005.

Tillema, M., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2011). Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning, 6*(3), 229–253. https://doi.org/10.1007/s11409-011-9072-x.

Torrance, M., Thomas, G. V., & Robinson, E. J. (2000). Individual differences in undergraduate essay-writing strategies: A longitudinal study. *Higher Education, 39*(2), 181–200.

Uto, M., Miyazawa, Y., Kato, Y., Nakajima, K., & Kuwata, H. (2020). Time-and Learner-Dependent Hidden Markov Model for Writing Process Analysis Using Keystroke Log Data. *International Journal of Artificial Intelligence in Education, 30*, 271–298. https://doi.org/10.1007/s40593-019-00189-9.

Van Waes, L., & Leijten, M. (2015). Fluency in Writing: A Multidimensional Perspective on Writing Fluency Applied to L1 and L2. *Computers and Composition, 38*, 79–95. https://doi.org/10.1016/j.compcom.2015.09.012.

Van Waes, L., Leijten, M., Mariën, P., & Engelborghs, S. (2017). Typing competencies in Alzheimer's disease: An exploration of copy tasks. *Computers in Human Behavior, 73*, 311–319. https://doi.org/10.1016/j.chb.2017.03.050.

Van Waes, L., & Schellens, P. J. (2003). Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics, 35*(6), 829–853. https://doi.org/10.1016/S0378-2166(02)00121-2.

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.

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing, 18*(6), 1499–1514. https://doi.org/10.1007/s00779-013-0751-2.

Wang, Y.-J., Shang, H.-F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26*(3), 234–257. https://doi.org/10.1080/09588221.2012.655300.

Wilson, J. (2017). Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing, 30*(4), 691–718. https://doi.org/10.1007/s11145-016-9695-z.

Wise, A. F., & Jung, Y. (2019). Teaching with Analytics: Towards a Situated Model of Instructional Decision-Making. *Journal of Learning Analytics, 6*(2), 53–69. https://doi.org/10.18608/jla.2019.62.4.

Woong Yun, G., & Park, S.-Y. (2011). Selective posting: Willingness to post a message online. *Journal of Computer-Mediated Communication, 16*(2), 201–227. https://doi.org/10.1111/j.1083-6101.2010.01533.x.

Xu, C. (2018). Understanding online revisions in L2 writing: A computer keystroke-log perspective. *System, 78*, 104–114. https://doi.org/10.1016/j.system.2018.08.007.

Zhang, M., Hao, J., Li, C., & Deane, P. (2016). Classification of Writing Patterns Using Keystroke Logs. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative Psychology Research: The 80th Annual Meeting of the Psychometric Society, Beijing, 2015* (pp. 299–314). Springer. https://doi.org/10.1007/978-3-319-38759-8_23.

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.), *IMPS 2017: Quantitative Psychology* (pp. 367–381). Springer International Publishing. https://doi.org/10.1007/978-3-030-01310-3_32.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.