Characterising Students' Writing Processes Using Temporal Keystroke Analysis

Donia Malekian[a]
malekiand@unimelb.edu.au

James Bailey[a]
baileyj@unimelb.edu.au

Gregor Kennedy[b]
gek@unimelb.edu.au

Paula de Barba[b]
paula.debarba@unimelb.edu.au

Sadia Nawaz[a]
nawazs@student.unimelb.edu.au

[a] School of Computing and Information Systems, [b] Melbourne Centre for the Study of Higher Education, University of Melbourne

ABSTRACT

This work aims to characterize students' writing processes using keystroke logs and understand how the extracted characteristics influence the text quality at specific moments of writing. Earlier works have proposed predictive models characterizing students' writing processes and mainly rely on distribution-based measures of pauses obtained from the overall keystroke logs. However, the effect of isolated phases of writing has not been evaluated in these models. Moreover, current theories on writing suggest that the quality of writing depends on when specific writing behaviours are performed. This view is not examined in the keystroke logging analysis literature. Addressing the mentioned challenges, the two contributions of this work are: a) characterizing students' writing processes connected to isolated writing phases and examining their influence on writing quality; and b) temporal analysis of keystrokes and investigating whether the significance of writing characteristics varies as students progress in their writing task. Our results suggest that characterizing students' writing based on isolated writing phases is slightly more predictive of writing quality. Additionally, the effect of several writing characteristics on writing quality changes when considering the time dimension.

Keywords

Writing process, Keystroke log, XGBoost, SHAP feature importance, Temporal analysis

1. INTRODUCTION

The recognized significance of the writing process has led to the emergence of a wide variety of research exploring the process of students' academic writing. The writing process, broadly including planning, writing and revising phases, is a non-linear process [10]. These phases often occur simultaneously, making it challenging for researchers to examine features related to specific writing phases.

One stream of writing research has focused on analyzing pauses in keystroke logs and associating them with different phases of the writing process [9]. This is often accomplished by exploring the distribution of pauses and then mapping the related parameters to specific writing processes. Although some keystroke log features can be a marker of a high writing quality, there is not always sufficient evidence for their relationship. This may be due to decisions made during data processing and analysis.

Donia Malekian, James Bailey, Gregor Kennedy, Paula de Barba and Sadia Nawaz "Characterising Students' Writing Processes Using Temporal Keystroke Analysis" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 354 - 359

Additionally, statistics and features collected by most studies examining the relationship between keystroke logs and the writing process mainly rely on aggregated or distribution-based representations of the overall keystroke logs [9]. Even when the data has been summarized to a high standard, neglecting the time dimension may hide the effect of specific behaviors at particular moments of the writing process. Temporal analysis has been suggested as a better way uncover the writing process and its related stages [3]. However, there are a few studies that combine writing process and temporal analysis, which mostly focused on think aloud procedures and offline measurements. These have been criticized as being inaccurate representations of the underlying writing process [16].

Therefore, our main aim is to examine students' writing processes and their relationship to writing quality using keystroke logs and temporal analysis. We use an innovative writing technology platform that assists in discriminating between writing phases, mainly planning and writing, by providing separate writing sections to students [17]. The temporal analysis provides insight about how the effect of each of those phases and keystroke log features on students' writing quality may change over time. This can support educators to make judgments regarding students' writing processes and change the ways they teach writing skills.

2. LITERATURE REVIEW

2.1 Writing Research

A common approach of writing research is to consider three phases to describe the writing process: pre-writing, writing, and post-writing [4]. Pre-writing is composed by planning the content of the text to be written. Writing is composing the ideas and transcribing them. Post-writing is revising or reviewing the written text or plan. For simplicity purposes, in our study these phases are referred to as planning, transcribing, and revising, respectively. Efforts in writing research have been made to identify behavioral features that could be an indication of these phases. Early writing research has heavily relied on self-report methods, such as think aloud protocols [3], to examine students' writing process. Over the recent decades, purpose-built software for writing research were developed which collected all information possible during the writing process. Initially these initiatives were restricted to laboratories, but recent advances in software development has now released such software naturalistic settings, allowing for researchers to examine the writing process in real educational environments [17].

2.2 Keystroke Logs in Writing Research

A key research area uses keystroke logging to characterize the writing process. Efforts here have focused on investigating the distributions of various kinds of pauses (e.g., inter-key, intra-word) and their relationship with writing quality. Among the existing studies, [7] found the exponentially modified Gaussian distribution a good fit for inter-key pauses and mapped specific pauses to

planning process in case they were identified to be long enough. In other work by [1], a mixture of lognormal distribution was used to describe pause pattern in students' inter-key and intra-word pauses. The parameters of the distribution were found to be correlated with writing score. Finally, [9] estimated the distribution of the inter-key and intra-word pause durations for each student and found the best fit to be a heavy-tailed probability distribution called a stable distribution. Because of nearly identical estimates for intra-word and inter-key pauses, they focused on only analyzing intra-word pauses. They found the estimated parameters for each student to be a strong predictor of final score utilizing a linear regression model.

Although these studies provide informative (predictive) models characterizing students' writing process, the effect of isolated phases of writing process have not been evaluated in these models. In addition, even though the use of simple models of regression and correlation analysis makes the interpretation of results easier, the use of more complex models may capture further information about the interrelationships between the extracted characteristics.

Another direction of research focused on using keystroke logs to more comprehensively characterize the phases comprising the writing process [9, 18]. In a study by [2] on modelling students' writing process, some measures were defined to model pauses, bursts and revisions. A burst is defined as the sequence of fast text production and can be identified based on the production of text between two pauses [2]. They suggest that long pauses may reflect planning, as the writers are more likely to have short but well-formed bursts of writing afterward.

Overall, extracted characteristics from keystroke logs in terms of burst, revision summaries and the pattern of pauses, provide important information regarding the underlying writing process. Association of each extracted characteristic with writing quality has been mainly considered as the metric for evaluating the usefulness of the feature [9, 18]. A major challenge is defining meaningful summaries from the writing keystrokes that represent a specific phase of students' writing process as much as possible. To address this problem, we use an innovative writing technology that more explicitly discriminates between the planning and writing phases, by providing separate writing sections to students [17].

Additionally, it is important to know not only which writing phases are relevant to successful writers, but when and in what order they engage with these phases. An approach to temporal data processing has been the aggregation of data in multiple consecutive episodes, so all participants have a similar number of observations [3]. They suggested that the relation between specific aspects of the process with writing quality varies as students' progress in their writing. For instance, the correlation of structuring activities and writing quality is highest at the start of the writing and is lower toward the end [3]. The importance of the temporal analysis of the writing process has been emphasized by several studies [16], however the data representation mostly relies on think aloud procedures and offline measurements that have been criticized as being inaccurate representation of the underlying writing process [16].

Although the importance of taking the moment(s) at which specific aspects of the writing process occur has been emphasized, this is not a dominant view in keystroke logging analysis.

2.3 Current Study

Our first aim is to characterize students' writing processes in terms of a set of features demonstrating the isolated phases. For this purpose, we focus on students' keystrokes characteristics while taking notes, writing the main body of the essay, and organizing references (each separately). Utilizing a machine learning model, we examine whether section specific characteristics are more predictive of writing quality compared to the characteristics extracted from the overall keystrokes. We also consider adding more features that estimate burst and revision behavior, as well as features representing the extent to which specific aspects of writing process were used during the writing. To evaluate how influential each feature is for each student's writing quality, we adopt a method called SHAP [13] that describes the importance of each feature on the model's prediction for each student. The second aim of the paper is to provide a temporal analysis which helps us to understand whether the importance of features varies over time or remains stable. We address this problem by breaking down students' keystrokes into several writing episodes and comparing the influence of each feature on writing quality across them.

Overall, in this study the following research questions are explored:

1. Do models that characterize the overall writing process miss informative features associating with particular phases of writing? Can we define new characteristics (features) from students' keystrokes to improve these models?
2. How predictive of writing quality are the extracted features at different times? Do we see evidence that the importance of features varies with time?

Our results highlight that characterizing students' keystrokes separately while writing, taking notes and organizing references is more predictive of their writing quality. Additionally, based on our findings, the importance of several features on writing quality changes over time. Investigation of the influential features for individual students clarified the non-linear relationship between some features and writing quality that confirms there exists considerable overlap and interaction between writing phases [4].

3. METHODS

3.1 Participants and Context

The study involves 107 students from the University of Melbourne who enrolled in a business undergraduate course. Students were asked to use a specific online word processing software called Cadmus to write a 1000-word essay as a part of their course, worth 10% of their final mark. Students had to choose between two topics and had 19 days to complete the essay. The performance was marked by teaching staff using a score between 0 to 100.

Cadmus has similar features to other word processing software tools such as body section for writing (body section), editing, highlighting, and additional features such as dedicated sections to take notes (note taking section), and to organize the reference materials (reference section) as well as a single paste restriction of 90 words. Cadmus records every keystroke in each section via the keyboard while students work on their assignment. A more detailed description of this software can be found in [17].

3.2 Data Processing

We next describe the procedure undertaken to characterize students' writing processes by engineering a set of features from keystroke logs. We also processed two concepts of *writing quality* and *writing episode* to assist on answering the research questions.

(A preliminary analysis revealed 4 students had less than 600 words in their essay. They were removed from further analysis.)

3.2.1 Pauses

In this study our focus is on inter-key pauses (the duration between successive keystrokes) that are more likely to be associated with

such processes as deliberation, text planning, reviewing the written text [9]. As [9] suggested, there is a tendency for inter-key pause durations to follow a stable distribution. They fitted this distribution to pause duration data to obtain an informative estimation of the related parameters for each student. We follow a similar procedure as [9]; however, we summarize pauses in each writing section of our dataset separately. This section-specific summarisation could reveal more explicit information regarding which processes were more likely to be engaged during the pause.

We aim to represent students' writing process while they are working on their essay, thus we decided to ignore the pauses more than 2 hours that meant the student had left the session.

In our dataset, the exploration of the distribution of pause durations for each student in each writing section reveals that there is a tendency for pause duration in each section to follow a heavy tail distribution, that means the majority of the pauses are short but there exist a few long pauses. Believing a stable distribution to be a plausible hypothesis, we fitted this distribution to each student's pause data to obtain an estimate of the related parameters in each section. This distribution needs four parameters (α, β, γ, δ) for the complete description.

- The parameter alpha $\alpha \in (0, 2]$, called the tail index. This parameter gives information about the height of the tails.
- The parameter beta $\beta \in [-1, 1]$, called the skewness parameter. The distribution is symmetric if $\beta = 0$. It is skewed to the right if $\beta > 0$, and to the left if $\beta < 0$.
- The parameter delta $\delta \in R$, is equal to median. Depending on how heavy the tail is, some extreme part of the data may need to be discarded to have a good estimate of this value.
- The parameter gamma $\gamma > 0$, called scale parameter is a measure of dispersion.

Parameters alpha and beta determine the distribution's shape, while parameters gamma and delta define the scale and location of it. For each student, in each section of our dataset we obtain these four estimated parameters. We also estimate these parameters for the overall keystrokes of each student (irrespective of the specific section) as suggested by [9] and consider them as a baseline for the purpose of evaluation.

3.2.2 Bursts

Burst summaries (i.e. mean length of the bursts) can reveal students' fluency in the transcribing phase of the writing process [18]. In keeping with the majority of the literature [14], we identify bursts by breaking up keystrokes at every pause that have longer than 2 seconds of inactivity. We apply this procedure in the body section of our dataset, where students write the main part of their essay. Considering the bursts in which at least one word is typed, we summarize burst length by two features based on the number of words in a burst: The *mean* and the *standard deviation of burst length* for each student. Additionally, we summarize burst duration by two features of *mean* and *standard deviation of burst duration.*

3.2.3 Revision

In this step, our aim is to isolate the revision phase of the writing process from the transcribing phase. Authors of [5] associated single backspaces to spelling correction which reflect self-monitoring, and multiple backspaces to editing in which longer revision occurred. Similarly, we summarize revision at small and large scale with a slight change in the definition; we identify revisions based on the number of word deletions in a writing burst rather than in isolation. We label a burst with single deletion as small, whereas bursts with multiple deletion as large-scale revision. Two features were extracted to provide measurements related to the revision phase of writing process*:* The *frequency of both small scale* and *large-scale revision bursts*.

3.2.4 Time percentage on each writing aspect

To summarize the extent of each specific aspect of writing that was used by each student, we introduce a new set of features, including *the percentage of the total writing time* dedicated to: note taking (total time in note taking section), small- and large-scale revisions (bursts of writing with small or large deletion), transcribing (total time of bursts), and reference organization (total time in reference section).

3.2.5 Writing quality

Writing quality corresponds to the students' final grades on the essay [9, 18]. Previous research has found that students' final grade may not be a reliable measure of success, due to variations in grading essay writing by raters [11]. To account for this variability, we map the students' writing quality to two categories - high and low level instead of the exact grade. In our dataset, the distribution of students' grades lies within the range 60 to 95. We adopt the median value as a threshold for this mapping which is 80. Based on this value we have 40 and 67 students having high quality and low quality writing respectively.

3.2.6 Writing episode definition

One approach to temporal data processing is the analysis of data in multiple episodes to ensure all participants have a similar number of observations [3]. In our study, students were asked to write a 1000-word essay, which provides a good criterion for defining the fixed observations. We define the episodes based on the keystrokes used to complete the fixed number of words in the essay. This way, we have meaningful episodes recording a specific draft of writing.

We split students' writing data into *n* writing episodes, $\{E_1 + E_2 + \ldots + E_n\}$, each of which records students' keystrokes used from the start of writing to the completion of *n*k* words of the essay. We define 5 writing episodes, each of which involves all the keystrokes from the start of writing to the completion of n*200 words. There is no theory for defining the number of episodes and the results may differ based on the choice of this number.

4. Data Analysis

We conducted two sets of analyses: one for each research question.

4.1.1 Research Question 1.

To answer whether models that characterize the overall writing process miss informative features associated with certain phases of writing, we compare the performance of a machine learning model trained on the pause related features extracted from overall keystrokes (*baseline* feature set) to the performance of a model trained on pause features of each writing section separately (*section-specific* feature set). Work in [9] also identified the *total time on task*, along with the pause related features as a strong predictor of writing quality. Thus, we include this feature in the *baseline* and *section-specific* feature sets. Our evaluation is based on the prediction performance of writing quality.

Then, we examine whether we can define new characteristics from students' keystrokes to improve our model. For this purpose, we evaluate our model by adding further features of burst and revision summaries (explained in section 3.2.2 and 3.2.3), as well as the features representing the extent to which specific aspects of the writing process were used (section 3.2.4). We refer to them as *combination* feature set. We utilised XGboost classifier [6] for the

prediction models. Even though this classifier generally provides a good prediction power compared to simple models of regression, understanding what the contribution of each feature were seems to be hard due to the complexity of the model. To evaluate and derive the influence of each feature on writing quality, we use SHAP (SHapley Additive exPlanations) algorithm which can be used to explain the output of any machine learning model [13]. In this algorithm the contribution of a feature is calculated by comparing what a model predicts with and without that feature. Every individual is assigned a SHAP value for each feature that determines the feature's contribution for a change in the model's prediction.

4.1.2 Research Question 2.

To examine how predictive of writing quality are the extracted features at different times, and how their contribution may vary, we broke down students' keystrokes into n episodes from each of which the predictive features (*combination* feature set) were extracted. Then for each episode the predictive power of the features on writing quality was evaluated using the XGBoost model. To identify and compare the contributing features to the models' prediction in each episode, SHAP algorithm was utilised.

5. RESULTS

5.1 Results for Research Question 1

First, we report and compare the prediction power of the introduced feature sets on writing quality. Then we derive and discuss the contribution of each feature on prediction.

5.1.1 Examining the prediction power of feature sets on writing quality

For each set of features a model was trained using leave one out cross validation (with 5-fold nested cross-validation for parameter optimization). The performance of each model on the prediction of writing quality is reported in Table 1 based on the metrics of accuracy and the area under the ROC curve (AUC) [12]. The accuracy and AUC were slightly higher for the *section-specific* feature set compared to the *baseline* set. This could indicate that characterizing the overall process irrespective of the isolated phases may miss specific moments where the features associated to certain writing phase become more important. Adding burst and revision summaries, as well as the time dedication features (*combination* feature set), to the model, improved the performance of the prediction significantly. The *combination* feature set obtained the highest prediction power implying that the introduced features were a better representation of the students' writing quality.

Table 1: Prediction power of each feature set on writing quality based on accuracy and AUC

Features set	Accuracy	AUC
Baseline	70.09	70.63
Section-specific	71.03	71.75
Combination	**81.31**	**82.72**

5.1.2 Examining the contribution of features on prediction

The next step was to evaluate the influence of each feature on predictions. For this purpose, we built an XGBoost model on the *combination* feature set (using 5-fold cross validation for parameter optimization). The model was then passed to the SHAP algorithm to explain the influence of each of features on the model's prediction for each student. Since we get individualized explanations of every feature for every student (based on the SHAP values), we can plot the distribution of the importance of each feature on the model's prediction, as is presented in Figure 1. The features are sorted by the mean of absolute SHAP values over all students to gain a global insight into the most influential features across all students. We observe the most influential features (globally) were total time on task *(TT)*, frequency of small revisions *(B-FSRev)*, and estimated gamma parameter from pauses in body section (*B-Pγ*) respectively, while some of the pause parameters estimated in the note taking and reference sections (i.e. *N-Pα, R-Pβ*) had the lowest contribution. For convenience of viewing, only the top-20 features that were globally influential are presented in the figure. In this figure, each row corresponds to a feature and every student has one dot on each row. The x position of the dot is the impact of that feature on the model's prediction for the student (SHAP value), and the color represents the value of that feature for the student (red high, blue low). This reveal, for example, that a high percentage of time on the reference section (*R-T*) increases the chance of having high quality writing for a subset of students (red dots in the *R-T* row, and on the right side of the plot).

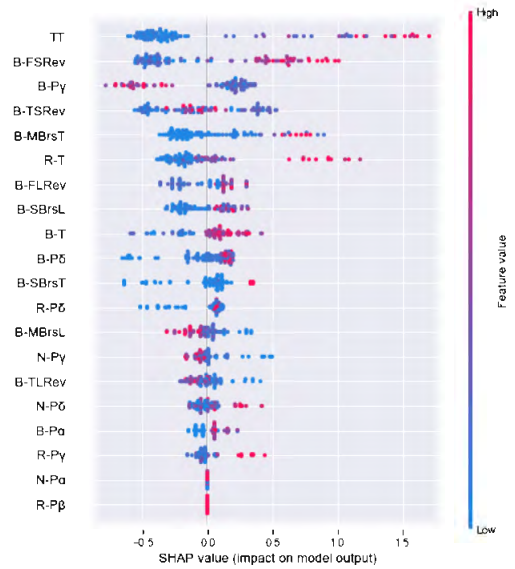

Figure 1: SHAP values for each feature indicating the influence of that feature on writing quality of each student. TT refers to the total time on writing. R, B, N before each feature name refer to Reference, Body and Note sections respectively. Pα, Pβ, Pδ, Pγ refers to pause parameters. TSRev, TLRev refers to the total time on large- and small- scale revisions respectively. FSRev, FLRev refers to the frequency of small- and large- scale revisions. T refers to the total time on specific section. MBrstT, SBrstT are mean and standard deviation of burst time. MBrsL, SBrsL refer to mean and standard deviation of burst length.

Below is our interpretation of some of the features detected as important by the prediction model. It is worth mentioning that this interpretation is intended for high-level model interpretation, and the model's decision making was more complex and took the interaction between features into account.

Total time on task *(TT)* was found to be the strongest predictor of writing quality. A subset of students (red dots on the related row of figure) who spent a lot more time than others on the essay are more likely to produce a high-quality writing. This supports previous research [9] and could be an indication of student's motivation in completing the writing task [10]. However, there are also students with lower time spent on task, but a higher chance of producing high-quality writing (blue dots on the right side of the plot in the

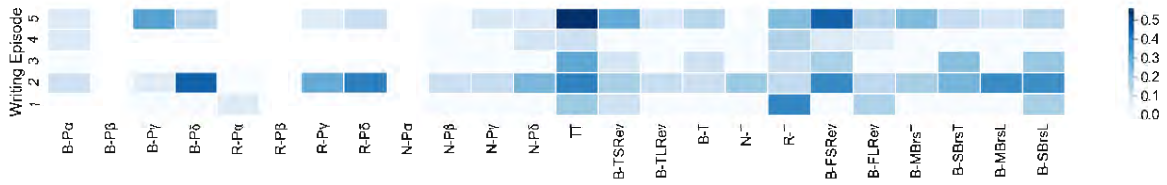

Figure 2: Visualization of the global importance of each feature on the prediction in each writing episode based on the SHAP values. Darker colors indicate higher contribution of that feature on prediction in that episode.

related row). This could indicate other features are impacting the contribution of this feature.

The coloring of the second most important predictor - the percentage of time on small revision (*B-FSRev)* shows us how a higher frequency of small revision means a higher the chance of producing high quality writing for most students. It could imply that the writing quality improves if revisions are performed more frequently. This is in agreement with previous research [15] suggesting that good writers stop their writing more often to perform revision and to correct their spelling and grammatical errors as they write compared to weak writers.

The estimated gamma and alpha parameters from pauses inside the body section (*B-Pγ, B-Pα*) shows a (negative, positive) association with the chance of producing a high-quality writing for most of the students. Together these parameters mean lower variation of pause durations without having very large pauses. This could be an indication of steadiness and fluency in writing and its direct relation with writing quality which is in agreement with previous research [9]. Percentage of time on reference section (R-T) also shows a positive effect on writing quality for several students. Again, we observe that this argument does not hold true for several students.

The higher percentage of time on transcribing (*B-T*) increase the chance of having high-quality writing. One alternative justification could be found in a study by [8]. They found that writers who spend most time on other aspects of writing such as planning tend to dislike writing and this may lessen the text quality.

Estimated delta and gamma parameters from pauses in the note taking section (*N-Pδ, N-Pγ)* shows a (positive, negative) association with the chance of producing a high-quality writing for most students. Together these parameters mean more large pauses and lower variation in pause duration. This means that steadily taking large pauses while taking notes leads to a higher chance of high-quality writing. Based on a work by [2], large pauses may reflect planning process. Since this pattern is observed in the note taking section of our dataset, we could more certainty connect these steady long pauses to thinking periods on planning.

Even though there are several features such as percentage of time on small revision (*B-TSRev*), mean burst length (B_MBrsL), estimated gamma from reference and body section (*R-Pδ*, *B-Pδ)*, that are detected as influential, we cannot observe their clear linear association with writing quality. For a subset of students, the higher value is associated with higher quality and for others it is the opposite. Again, this indicates that the importance of these features is impacted by other features.

5.2 Results for Research Question 2

In this section we answer the second research question in two steps.

5.2.1 Examining the predictive power of features on writing quality over time

Using the XGBoost classification algorithm, we examined the predictive power of the extracted feature in our study (*combination* feature set) on writing quality at each writing episode. We report our result using the metrics of accuracy and area under the ROC curve (AUC), based on leave one out cross-validation (with 5-fold hyper parameter optimisation). The outcome is shown in Table 2 demonstrating a high and, in some episodes, moderate success rate in predicting students' writing quality (better than random chance).

Table 2: Prediction power of extracted features in each writing episode based on the accuracy and AUC

Episode#	Accuracy	AUC
1	69.16	70.26
2	67.29	67.12
3	69.16	66.01
4	79.43	79.22
5	81.31	82.72

5.2.2 Examining the contribution of features on prediction over writing episodes

Our next step was to examine the contribution of each feature on the prediction of writing quality at each writing episode and examine whether the contribution varied. For this purpose, an XGBoost model was trained based on the *combination* feature set in each episode (using 5-fold cross validation for parameter optimization). Each model was then passed to the SHAP algorithm to explain the importance of each of feature on the prediction. The result is reported in Figure 2, in which the global contribution of each feature in each writing episode is visualized based on the mean of absolute SHAP values for that feature over all students.

A darker color indicates a higher contribution of that feature on the prediction model in that episode. We see the contribution of features on writing quality are quite different across the writing episodes. Although the importance of total time on task *(TT)* is relatively stable over all the episodes, the importance of the note taking related features such as percentage of time on note taking section *(N-T)*, and estimated pause parameters *(N-Pδ, N-Pβ)* were found to be more influential in the 2nd writing episode. The patterns of pauses in the reference section (i.e. percentage of time on the reference section *(R-T)*, as well as the estimated gamma and delta parameters (*R-Pγ*, *R-Pδ*) also show stronger influence on the prediction of writing quality at the beginning of writing.

Our temporal analysis reveals that the importance of writing behaviour on writing quality may change over time. Thus, characterizing students' writing process irrespective of time may hide the effect of meaningful and predictive writing behaviors. Moreover, this analysis could be used to predict students writing quality over time and act as a filter for early targeting of students with different writing qualities opening up feedback opportunities.

6. DISCUSSION

This study was based on a fully online web authoring tool called Cadmus. The availability of separate sections for body text, note taking and referencing allowed us to separately extract the

keystroke logs of different sections. From these logs we were able to engineer multiple sets of features capturing different aspects of students' writing processes ranging from patterns of pauses, burst and revision summaries as well as the time dedicated to specific aspects of the writing process. We compared the performance of a model trained on the pause pattern related features extracted from overall keystrokes (*baseline*), to the performance of a model trained on pause pattern related features of each writing section separately. The section-specific model performed slightly better. This indicates that the baseline model may miss or "overlook" on specific moments where the features associated with certain writing phases become more important. The performance of the section-specific model was further improved by adding more features including burst and revision summaries and percentage of time on specific activities.

The feature importance of the resultant model is visualized for every student showing how specific behaviour that have positive effect on writing quality for one student may have negative effect for another because of the impact of other features. This is consistent with a theory of writing which suggests there exists considerable interaction and overlap between writing phases [4]. This also emphasises the necessity of using models that capture the interrelationship between features rather than simple correlation and regression analysis.

The current study also examined whether the influence of extracted features on writing quality varies during specific moments of writing. Based on our results, the influence of several features varied across writing episodes indicating the importance of taking the temporal aspects of writing process into account. Through this study, we hope to develop a better understanding of students' writing process in authentic educational settings.

More detailed results, discussions, and additional figures that could not be included in this version of the paper for the reason of space, is available in the longer version of the paper on this link.

6.1 Limitations and Future Work

The first limitation of this study is the generalizability of the interpretations regarding the influential features. This study was based on the essay writing of undergraduate students with diverse writing and language backgrounds. The influential features might differ for data from a more diverse set of students and across variations in topic, genre and prompts. The next limitation arises from considering all the activities in the note taking and reference sections to be associated with the related phase of writing. This association is irrespective of the actual written text. For the next study we aim to consider the actual text entered in each section. This may help to distinguish between weak and strong planning behaviour and the related impact on writing quality.

REFERENCES

[1] Almond, R., Deane, P., Quinlan, T., Wagner, M. and Sydorenko, T. 2012. A preliminary analysis of keystroke log data from a timed writing task. *(ResearchReportNo.RR-12-23).* (Dec. 2012), Princeton, NJ: Educational Testing Service. DOI:https://doi.org/10.1002/j.2333-8504.2012.tb02305.x.

[2] Baaijen, V.M., Galbraith, D. and de Glopper, K. 2012. Keystroke analysis: reflections on procedures and measures. *Written Communication*. 29, 3 (Jul. 2012), 246–277. DOI:https://doi.org/10.1177/0741088312451108.

[3] den Bergh, H. and Rijlaarsdam, G. 1996. The dynamics of composing: modeling writing process data. *The Science of Writing: Theories, Methods, Individual Differences, and Applications*. New York: Lawrence Erlbaum Ass. 207–232.

[4] Biggs, J. 1988. The role of metacognition in enhancing learning. *Australian Journal of Education*. 32, 2 (Aug. 1988), 127–138. DOI:https://doi.org/10.1177/000494418803200201.

[5] Chanquoy, L. 2009. Revision processes. *The SAGE Handbook of Writing Development*. London: SAGE Publications Ltd. 80–97. DOI:10.4135/9780857021069.n6.

[6] Chen, T. and Guestrin, C. 2016. XGBoost: a scalable tree boosting system. *Proc. of the 22nd Conference on Knowledge Discovery and Data Mining* (New York, USA, Aug. 2016), 785–794. DOI:10.1145/2939672.2939785.

[7] Chukharev-Hudilainen, E. 2014. Pauses in spontaneous written communication: a keystroke logging study. *Journal of Writing Research*. 6, 1 (Jun. 2014), 61–84. DOI:https://doi.org/10.17239/jowr-2014.06.01.3.

[8] Green, D.W. and Wason, P.C. 1982. Notes on the Psychology of Writing. *Human Relations*. 35, 1 (Jan. 1982), 47–56. DOI:https://doi.org/10.1177/001872678203500104.

[9] Guo, H., Deane, P.D., van Rijn, P.W., Zhang, M. and Bennett, R.E. 2018. Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*. 55, 2 (Jun. 2018), 194–216. DOI:https://doi.org/10.1111/jedm.12172.

[10] Hayes, J. and Gradwohl Nash, J. 1996. On the nature of planning in writing. *The Science of Writing:Theories, Methods, Individual Differences, and Applications.* Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc. 29–55.

[11] Kayapinar, U. 2014. Measuring essay assessment: intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research*. 14, 57 (Oct. 2014), 113–135. DOI:https://doi.org/10.14689/ejer.2014.57.2.

[12] Ling, C.X., Huang, J. and Zhang, H. 2003. AUC: a statistically consistent and more discriminating measure than accuracy. *Proc. of the 18th Conference on Artificial Intelligence* (San Francisco, CA, USA, Aug. 2003), 519–524.

[13] Lundberg, S.M. and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30* (Long Beach, Ca, US, Dec. 2017), 4765–4774.

[14] Rosenqvist, S. 2015. *Developing pause thresholds for keystroke logging analysis*. B.A. thesis. University of Umea, Sweden. Retrieved from http://umu.diva-portal.org/smash/get/diva2:834468/FULLTEXT01.pdf.

[15] Stallard K.C. 1974. An analysis of the writing behavior of good student writers. *Research in the Teaching of English*. 8, 2 (Summer. 1974), 206–218.

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

[17] Trezise, K., de Barba, P.G., Jennens, D., Zarebski, A., Russo, R. and Kennedy, G. 2017. A learning analytics view of students' use of self-regulation strategies for essay writing. *Proc. of the ASCILITE 2017* (Toowoomba, QLD, AUS, Nov. 2017), 411.

[18] Zhang, M., Hao, J., Deane, P. and Li, C. 2018. Defining personalized writing burst measures of translation using keystroke logs. *Proc. of the 11th Conference on Educational Data Mining* (Buffalo NY, Jul. 2018), 549–552.