Talebinamvar and Zarrabi *Language Testing in Asia* (2022) 12:6
https://doi.org/10.1186/s40468-021-00150-5

Language Testing in Asia

RESEARCH Open Access

Clustering students' writing behaviors using keystroke logging: a learning analytic approach in EFL writing

Mobina Talebinamvar[2] and Forooq Zarrabi[1*]

* Correspondence: forooq.zarrabi@ustmb.ac.ir
[1]Engineering college, Alzahra University, Vanak Village Street, Tehran 19938 93973, Iran
Full list of author information is available at the end of the article

Abstract

Feedback is an essential component of learning environments. However, providing feedback in populated classes can be challenging for teachers. On the one hand, it is unlikely that a single kind of feedback works for all students considering the heterogeneous nature of their needs. On the other hand, delivering personalized feedback is infeasible and time-consuming. Available automated feedback systems have helped solve the problem to some extent. However, they can provide personalized feedback only after a draft is submitted. To help struggling students *during* the writing process, we can use machine learning to cluster students who benefit the same from feedback using keystroke logs. We can apply the results in automated feedback systems that provide process feedback. In this study, we aim to find homogeneous student profiles based on their writing process indicators. We use fourteen process indicators to find clusters in the data set. We used these measures in a four-stage analysis, including (a) data preprocessing, (b) dimensionality reduction, (c) clustering, and (d) the analysis of the writing quality. Clustering techniques identified five different profiles: Strategic planners, Rapid writers, Emerging planners, Average writers, and Low-performing writers. We further validated the emerged profiles by comparing them concerning students' writing quality. The present work broadens our knowledge of how students interact with writing tasks and addresses how variations in writing behaviors lead to qualitatively different products. We discuss the theoretical underpinnings and potentials of finding profiles of students during writing in higher education.

Keywords: Clustering, EFL writing, Keystroke logging, Machine learning, Process indicators

Introduction

Feedback is believed to have a significant role in learning (Lam, 2021) and is the main factor in improving writing quality (Wilson et al., 2021; Sarid et al., 2021). To assist teachers in providing frequent and timely feedback to large numbers of students, various automated writing evaluation (AWE) tools (Ranalli, 2021) have been developed to provide automated written corrective feedback (Ranalli, 2019) on students' writing. However, these systems put attention to giving feedback on "the writing product"

(Zhang & Deane, 2015; Vandermeulen et al., 2020) rather than "the writing process" (Allen et al., 2016). Though the design of technology-enhanced forward-looking feedback (Cunningham, 2019a, 2019b; Cunningham & Link, 2021; Saricaoglu, 2018; Tseng & Yeh, 2019) on learners' product is an achievement over more traditional forms of feedback that only justify the grade, if it only "points to future performance" (William, 2010), it might not assist students who struggle *during* the writing process. As a result, the chances of engagement with the feedback decreases (Otnes & Solheim, 2019). The importance of timely feedback and providing opportunities to act on it in the learning and assessment cycle is also stated in the literature (Carless et al., 2011; Winstone & Boud, 2020). According to Lam (2021), outcome-related feedback does not leave any room for learners to improve their performance based on the feedback. When the feedback is provided too late to be used, the students may value it less. Subsequently, they may get less motivated to engage with the feedback (Steen-Utheim & Hopfenbeck, 2019; Henderson et al., 2019). As a result, the lack of motivation might tremendously affect the results of the feedback.

Due to the limitations of output-related feedback, we witness a growing interest in researching process-oriented feedback to inform automated feedback systems. However, designing process-oriented feedback systems is not without challenge. One main concern is the heterogeneous nature of the students' profiles. On the one hand, it is improbable that a single kind of feedback can be ideal for all students. On the other hand, delivering personalized feedback is not feasible since students' characteristics are highly variable. As a result, it is challenging to identify the right kind of feedback. To address this challenge, researchers in many fields of education have adopted clustering, or cluster analysis, which is an unsupervised machine learning (ML) task. It involves automatically discovering natural grouping in data (Jain, 2010). In writing courses, clustering finds groups of students with similar writing behavior. Then, teachers can provide more accurate, personalized feedback to students and can help them in optimizing their cognitive potential (Conijn et al., 2020; Kochmar et al., 2020; Zheng et al., 2021). Consequently, they can improve self-regulated learning (Chou & Zou, 2020). At a higher level, clustering students' writing behavior can improve automatic assessments (Brooks et al., 2014; Zehner et al., 2016), especially in high-stake tests like TOEFL and IELTS. To date, many studies have clustered students into meaningful groups to inform interventions (Hung et al., 2015; Mojarad et al., 2018). Extending these lines of research to keystroke logging (KL) studies, in the present work, our goal is to identify clusters of certain manifest writing behaviors to inform process feedback tools. Previous studies have attempted to find writing processes that predict high-quality writing products (Choi & Deane, 2021; Sinharay et al., 2019). However, the nature of students' profiles is affected by several intervening factors, and this heterogeneous nature does not allow for generalizing specific effects beyond a specific writing session or task (Conijn et al., 2019). To address this issue, we use clustering techniques that automatically discover groups of students with the same writing behaviors. Then, feedback adapted to the needs of each profile can be delivered to them. As a result, clustering can be a more efficient approach in feedback studies compared with other approaches employed so far. In addition, we utilize KL features for profiling students, which can capture and display the writing processes involved in each student profile. This information can be used to address the process aspect of feedback more precisely using a multi-stage feedback

design (Carless & Boud, 2018). Moreover, we aim to understand how these profiles and KL features relate to writing quality. This relationship was overlooked in previous studies. Therefore, in the current study, we use ML techniques to discover writing profiles by using KL data and the writing quality of English as a foreign language (EFL) students.

Research questions

Taking previous studies into considerations, we address the following research questions:

1. Are distinct clusters revealed based on process indicators produced during argumentative writing, and if distinct clusters are revealed, what KL features discriminate students in each cluster?
2. How do discriminating KL features in each cluster relate to the quality of their final product?

Related work

Providing process-oriented feedback is a challenge faced by AWE tools designers. We suggest using KL in designing these tools, and as a first step, we focus on identifying profiles of students. Below, we present an overview of KL studies. Then, we review studies that (a) use ML techniques to analyze students' writing behaviors, (b) primarily focus on writing quality, or (c) use KL features for profiling students.

An overview of KL studies

KL is an unobtrusive approach to get fine-grained process data on every keystroke during writing (Vandermeulen et al., 2020; Guo et al., 2019). Extensive empirical research with different orientations has been done on writing to discover the relation between KLs and underlying cognitive writing processes (Leijten & Van Waes, 2013). Some researchers attempted to develop models based on KLs to give teachers predictive power over students' writing performance (Choi & Deane, 2021; Zarrabi & Bozorgian, 2020). The relationship between writing processes and writing quality has also been explored in several studies. Except for a few recent studies (Sinharay et al., 2019; Choi & Deane, 2021), these correlational studies mainly were engaged in single unit measures of writing behavior such as total time on task (Sinharay et al., 2019; Zhang et al., 2019), pause time (Chukharev-Hudilainen et al., 2019; Révész et al., 2019), and revision (Al-Saadi & Galbraith, 2020; Bowen & Van Waes, 2020) using various analysis methods (Sinharay et al., 2019; Zarrabi & Bozorgian, 2020; Wallot & Grabowski; 2019; Conijn et al., 2020). Findings of these theory-driven KL studies indicate that researchers have come to a consensus about the association of some process indicators with some underlying cognitive processes. For instance, some studies report that burst length, in-word typing speed, between-word pause length, and initial pause time before typing a word are indicators of writing proficiency (Zhang & Deane, 2015). Other features, however, were only suggested to relate to some higher-order cognitive processes (Choi & Deane, 2021; Barkaoui, 2019; Baaijen & Galbraith, 2018). These findings might be used for human feedback to some extent (Ranalli & Yamashita, 2020; Vandermeulen et al., 2020), but

they are not applicable for designing precise automated feedback systems that affect large populations. In addition, They are all product-oriented studies, and thus, their findings can hardly, if at all, be employed in process-oriented feedback tools.

Research works aimed at predicting writing quality from KL

Several recent studies have adopted a data-driven approach to inform designing precise AWE tools. They focused on developing models that could automatically predict writing scores using KL. Using boosting and linear regression, Sinharay et al. (2019) analyzed 38 KL features from two persuasive essays (sample sizes of 825 and 832 for each task) to predict essay scores. The results indicated the process features predict the essay scores (*RMSE* = 0.50 on a scale of 1–5) only slightly worse than the product features (*RMSE* = 0.44). Burst length, number of bursts, typing speed, and time on task had the highest predictive power. Moreover, Sinharay et al. (2019) reported that boosting predicted essay scores slightly better linear regression. Likewise, Choi and Deane (2021) analyzed 956 log files from 576 adult EFL learners in two source-based writing tasks and two argumentative essays using an exhaustive subset search and the LASSO approach. They reported that two to five features could predict the writing quality on different tasks (*PRMSE* = .29 to .48). The number of keystrokes had the highest predictive power across all tasks. Lastly, Conijn et al. (2021) analyzed 54 keystroke features from 126 students performing a timed academic summarization task. Regression and classification models were used to predict final scores at three-time points during the writing process. The classification models were pretty better than the class baseline (highest *AUC* = 0.57), and the regression models were even below. Furthermore, the relationship between the KL features and writing quality was not stable during the writing process. In contrast with previous studies, Conijn et al. (2021) reported a less-than-obvious relationship between keystroke features and writing quality. In sum, fluctuations in selecting keystroke features, sample sizes, and method of analysis and contradictory results in these studies make it challenging to determine which features are most relevant for predicting writing quality. As a result, other research studies employed the technique to address this shortcoming. They grouped KL features into meaningful clusters to be used for interventions.

Research works aimed at clustering KL

The advent of keystroke logging in writing research has more accurately measured certain aspects of the writing process, and their relation to writing quality. For instance, total time on task (Zarrabi & Bozorgian, 2020; Sinharay et al., 2019; Zhang et al., 2019) and the total number of words (Allen et al., 2016; Likens et al., 2017) have positive associations with writing scores in several studies. However, they explain the different amounts of variance for the writing quality in different studies. Fluctuations of variance explained by each of these features are understandable, given that there is no "single" writing process that produces a high quality for every student. As a result, clustering techniques can provide more valid results by grouping features automatically and showing how these profiles relate to writing quality.

Zhang et al. (2017) used clustering techniques to group students based on four fundamental writing performance indicators. They identified four distinct clusters of

writers. The four revealed clusters displayed different sequential patterns throughout writing on the mean essay score, mean total time on task, and number of words in their final drafts. Although the work intends to inform corrective feedback, the number of process indicators used might not be sufficient for identifying at-risk students from high achievers, mainly because the indicators of pausing behavior are missing. In addition, the emerged profiles need to be validated by comparing them to students' final product.

In more recent work, Shen and Chen (2021) profiled eight Chinese EFL learners' pausing behavior across writing skill levels at word and sentence level pause indicators. Findings from the qualitative and quantitative analysis revealed that overall, there was not a significant difference between more-skilled writers and less-skilled writers on the total pause time. However, the two groups showed very different pausing strategies: less-skilled writers paused more at word boundaries, but more-skilled writers employed more strategic pauses, i.e., paused at sentence boundaries. This result contrasts previous studies' findings (Conijn et al., 2021; Medimorec & Risko, 2017) which report that pauses between words are related to planning and higher writing quality. Though Shen and Chen (2021) provide some insight into which keystroke features relate to writing quality, their sample size is not large enough to guarantee substantial segments. Automated ML techniques, large samples, and a wide variety of KL features for clustering are needed to inform feedback on the quality of writing.

Methods

Instruments

We used an argumentative writing task the data using Inputlog 8.0.0.1 (Leijten & Van Waes, 2013). The students had to provide enough facts, evidence, and warrants to support their claim for/against a controversial issue (free state funding education). We used two different rubrics for scoring students' essays: TOEFL iBT independent writing rubrics to assess the overall quality of writing. The other one examined the quality of the arguments based on Argumentative Essay Rubrics (Appendix). All essays were rated against each rubric. Two raters double-scored each essay. All essay scores are first estimated as the sum of the two rubric scores (each rubric scales from 0 to 15) for each essay. Then, the average of the two human-rater scores is computed and considered the total score. Following Zhang et al. (2017), the inter-rater reliability (.71 and .73) was measured using quadratically weighted kappa for the two rubrics.

Participants

The data collected are part of a research project involving 20 classes at Mazandaran University of Science and Technology (MUST). A total of 180 male and 458 female Iranian B.A. students (mean age = 19.3 years, *SD* = .47; year 3; native language = Persian) participated in this study. Two participants did not follow the directions and made a scratch file first and then pasted the data from the original file into Inputlog file. Four students wrote very concise, off-topic essays. One student's log file was erroneous and could not be processed. Excluding these participants, 631 files remained for us to process.

Data set

We used Inputlog 8.00.1 for data collection. After removing erroneous log files, meaningless, or off-topic essays, our error-free dataset included 631 files submitted by the participants. The clustering analyses focused on KL features from previous research that predict writing quality. Additionally, we added some features which correlated with writing quality in our unpublished paper. Then, clustering analyses were conducted to automatically discover which KL features cluster together, considering the final score. The result of the Euclidean distance in *K*-means clustering indicated that only fourteen KL features could form clusters. These variables can help us gain information about associated but distinct sides of writing such as pausing behavior, fluency, and time on task, and the final product (Table 1).

All the indicators presented in Table 1 are straightforward and automatically extracted from Inputlog 8.00.1. Three indicators, however, were extracted semi-automatically and need to be defined: the longest pause, location of the longest pause, pause variance. We define them as (1) the longest pause between two sequential keystrokes, (2) the longest pause location indicator defined as the number of the interval for the longest pause between two sequential keystrokes during the writing, and (3) the variance indicator defined as the variation in the length of pauses between two sequential keystrokes during the whole writing process respectively. We included the longest pause and its two related indicators as they are believed to depict strategic processes like discourse-level and sentence-level planning (Zhang et al., 2017). Apart from the above process indicators, we had the total score, which shows the general writing quality of each student.

Data analysis

To address the first research question, we inspected possible student profiles based on KL indicators using clustering. To this end, we implemented a four-stage process, namely: data preprocessing, principal component analysis (PCA), clustering (*K*-means), and comparing clusters and writing quality (Fig 1). First, we needed to address the diversity of measurement units for process indicators. Process indicators had various units of measurement (e.g., second for pausing behaviors or character for mean typed in p-burst). Therefore, at the data preprocessing stage, we normalized the data using *Z*-score normalization. At the second stage, we addressed the issue of the possible correlation between the process indicators. To transform probably correlated indicators into fewer uncorrelated indicators, we used principal component analysis (PCA). PCA is

Table 1 Process indicators and their related aspects of writing

Pausing behavior	The longest pause	Location of the longest pause	Pause variance	Geometric mean of before-sentence pauses	Total pause time (in seconds)
Fluency	Median typed in p-bursts (characters)	Total typed per minute (including spaces)	Geometric mean of within-word pauses	Geometric mean of before-word pauses	
Time on task and the final product	Total words in main document	Mean word length	Total keystrokes	Total active writing time (in seconds)	Total process time (in seconds)

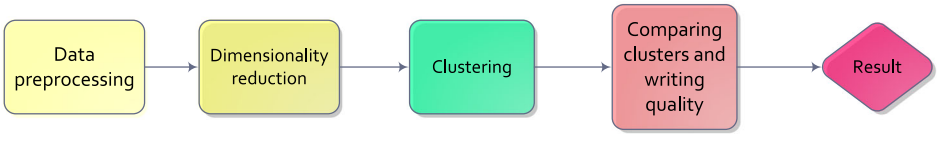

Fig. 1 The flowchart of data analysis

used for data dimensionality reduction. It substitutes the higher dimensional primary data with fewer non-correlated extracted vectors ("principal components").

Next, we decided on the optimal K (number of clusters) (Fig. 2). To this end, we employed the Elbow method and the Silhouette method to identify the number of clusters (k). Besides the methods mentioned above, we determined the number of clusters using Hierarchal clustering, which is an algorithm that builds a hierarchy of clusters. These methods are not alternatives to each other for determining the best K for K-means. Instead, we used them together for a more confident decision.

Finally, when we determined the number of clusters, we used the clustering algorithm to identify clusters with similar process indicators. As the K-means clustering algorithm is easily interpretable, it is widely used for data clustering, and thus, we used it to discover the actual student groups. Then, to name the identified clusters, we used box plots to visualize the clusters. When student clusters were identified, we used the Mann-Whitney U test to examine the significance of the difference among clusters on the process indicators. The Mann-Whitney U test was selected since it is a nonparametric statistical method and does not make any assumption about data distribution. Besides the process indicators, each student had a score for their writing quality. Thus, we compared the writing quality among various clusters.

We explored if the difference in students' writing quality was significant by employing the Mann-Whitney U test. We used the Python programming language (Pandas, Numpy, Matplotlib, Scikit-learn, Yellowbrick, Statistics, Scipy) in the Jupyter notebook application to analyze the data. The detailed clustering algorithm is shown in Fig. 3.

Results

Principal component analysis

PCA displayed four principal components (PCs), presented in Table 2. Table 2 illustrates these factor weightings and the variance proportion described by each component. In total, these PCs explain around 77% of the data variance. Thus, they make the input to hierarchical clustering and k-means. These PCs strongly refer to all process indicators, and consequently, the selection of process indicators is established as well.

Clustering

Using Elbow, Silhouette, and Hierarchical methods, we identified five clusters (k = 5). As illustrated in the Elbow method plot (Fig. 4) and the Silhouette method plot (Fig. 3), the curve displays k = 5 clusters.

The average Silhouette score (.4178) for k = 5 further confirms five potential clusters in the data set. Each color in Fig. 5 represents the distribution of the silhouette score of

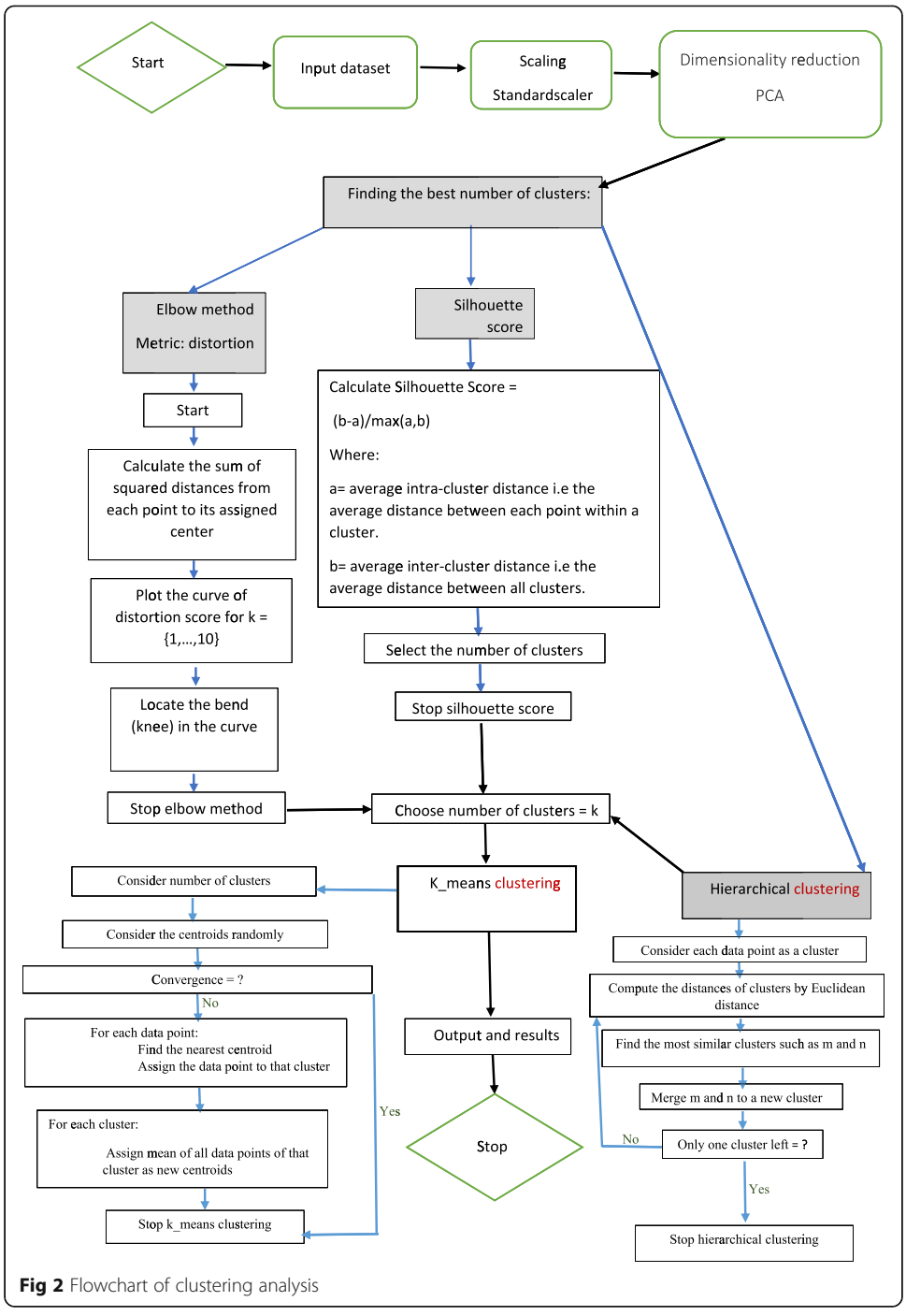

Fig 2 Flowchart of clustering analysis

each student that falls within the cluster. The graph for a cluster gets wider when the number of students within that cluster increases.

The Hierarchical method also revealed $k = 5$ clusters. In the Hierarchical method plot, we cut the dendrogram so that the tallest vertical line is cut. This plot displays $K = 5$ clusters (Fig 6).

To find distinct student profiles, we employed K-means clustering for $k = 5$ clusters and the Euclidean distance as the distance measure. To this end, the data from the first

```
Begin
    Read dataset
    Scaling
    Dimensionality reduction
    Finding the best number of clusters:
            Elbow method:
                Calculate the sum of squared distances from each point to its assigned center
                    Plot the curve of distortion score for k = { 1 , … , 10 }
                    Locate the bend (knee) in the curve
                    End elbow method
            Silhouette score
                    Calculate Silhouette Score = (b-a)/max(a,b)
                        Where:
                            a= average intra-cluster distance
                                i.e the average distance between each point within a
cluster.
                            b= average inter-cluster distance
                                i.e the average distance between all clusters.
                     Select the number of clusters
                     End silhouette score

            Hierarchical clustering
                1. Consider each data point x as a cluster ; (k = x)
                2. Compute the distances of clusters by Euclidean distance
                3. Find the most similar clusters such as m and n
                4. Merge m and n to a new cluster
                5. While (number of clusters !== 1) :
                                                    Go to step 2
                6. End hierarchical clustering

    Number of clusters = k
    K_means clustering
            Consider the centroids randomly
            While (k_means have not converge)
            For each data point:
                Find the nearest centroid
                        Assign the data point to that cluster
                  For each cluster:
                        Assign mean of all data points of that cluster as new centroids
             End k_means clustering
End
```

Fig. 3 Clustering algorithm

4 PCs were used. The mean values of each process indicator for the discovered clusters are presented in Table 3. We compared clusters based on these values to interpret these clusters.

Subsequently, the Mann-Whitney *U* test is run to determine if the difference of each process indicator is statistically significant among clusters. The differences are presented in Table 4. As illustrated in Table 4, each pair of clusters is different in all process indicators. The differences are significant at $p < 0.05$. This finding suggests that the proposed behavioral indicators can be used to name and describe the identified clusters.

In the following section, the identified clusters are named and described in detail based on statistics in Tables 3 and 4. To describe clusters, boxplots of process indicators for each cluster are used to present clusters in a more human-readable way.

Table 2 Process indicators weights and the explained proportion of variance for PCs

Process indicators	PC 1	PC 2	PC 3	PC 4
The geometric mean of within-word pauses	0.052	0.10	− 0.09	0.75
The geometric mean of before-word pauses	0.32	0.19	− 0.14	0.23
The geometric mean of before-sentence pauses	0.22	− 0.20	0.03	0.27
The longest pause	0.28	0.19	0.46	− 0.12
Location of the longest pause	− 0.12	− 0.29	− 0.09	0.28
Pause variance	0.41	0.02	0.34	− 0.06
Total pause time	0.36	0.30	− 0.15	− 0.008
Total active writing time	− 0.20	0.36	− 0.29	− 0.10
Total process time	0.27	0.38	− 0.22	− 0.03
Total words in main document	− 0.26	0.42	− 0.02	0.10
Mean word length	0.09	− 0.07	− 0.10	0.29
Mean typed in p-bursts	− 0.10	0.15	0.62	0.27
Total typed per minute	− 0.43	0.10	0.22	0.10
Total keystrokes	− 0.22	0.44	0.08	0.04
Explained variance proportion	0.30	0.26	0.11	0.09

Student profile names

Considering previous studies' findings, we expected to find statistically distinct student profiles. More specifically, we expected to find three main classes of proficient, Average, and low-performing writers. The names and description profiles are as follows:

1. Strategic planners (cluster 4): They put in the high effort, write fluently, and start writing with a relatively long pause followed by pauses of different lengths distributed throughout the writing process.
2. Rapid writers: (cluster 3): these students are fluent writers, showing persistence in their writing effort. They start writing with no pause and will not stop until later

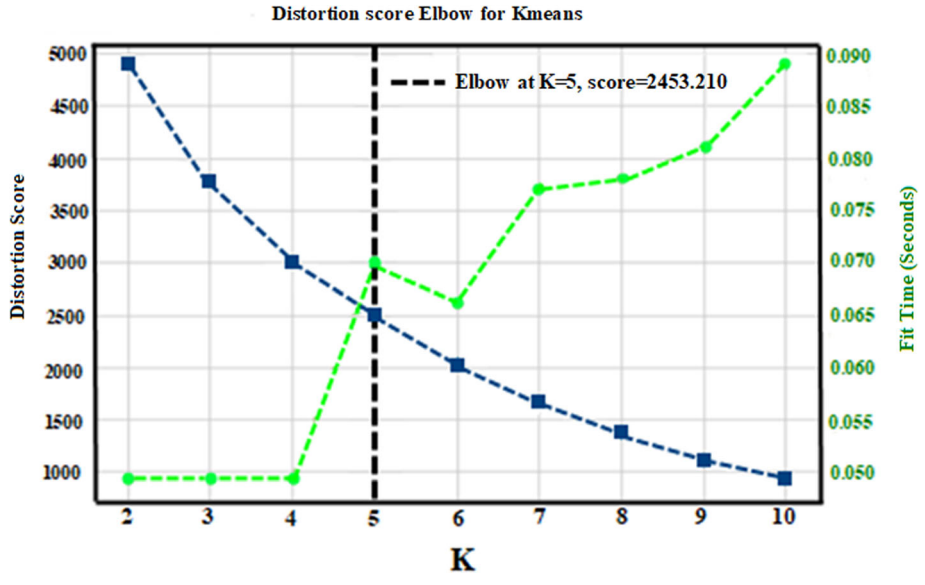

Fig. 4 The Elbow method plot

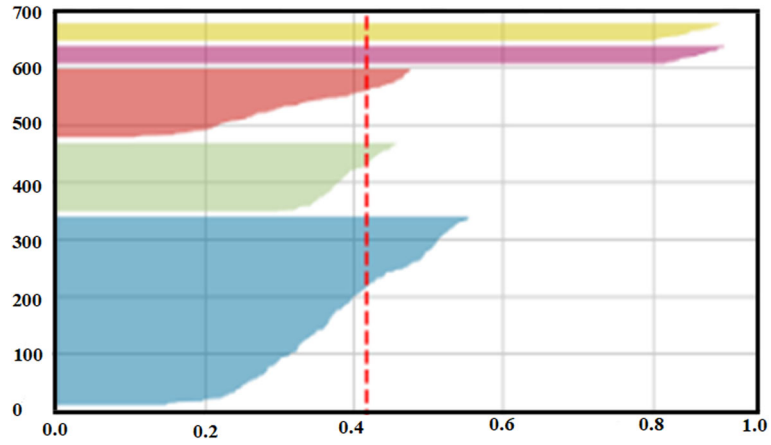

Fig. 5 The silhouette score

stages in writing, where they still keep the longest pause as short as possible. Varied-length pauses are distributed all through their writing process. Their total pause time is evidently below average, and according to time indicators, they finished the task faster than others did.

3. Emerging planners (cluster 1): They show high writing fluency and acceptable effort in writing. Their writing is characterized by a very long pause at the initial stages of writing followed by relatively same-length pauses.
4. Average writers (cluster 2): This group shows the mean value in all process indicators. These students reveal a medial position compared with students in other profiles.
5. Low-performing writers (cluster 5): This group puts in a low effort, low fluency. They demonstrate below-average pausing behaviors.

The identified clusters are interpreted using variables associated with three aspects of the writing process, i.e., pausing behavior, fluency, time on task, and final product. These three aspects and their related variables are presented in Table 1. The clusters' names are assigned based on behavioral indicators of each cluster.

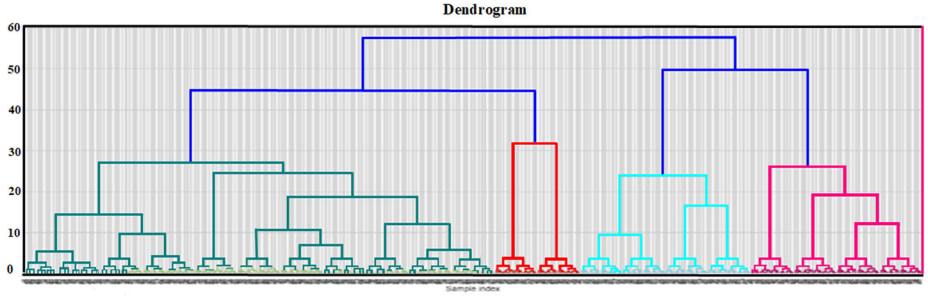

Fig. 6 Dendrogram of hierarchical clustering

Table 3 Average values of process indicators for each cluster

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
The geometric mean of within-word pauses	4.34	4.95	5.22	5.27	4.74
The geometric mean of before-word pauses	6.09	7.65	6.006	7.13	5.65
The geometric mean of before-sentence pauses	15.75	32.82	16.64	9.79	37.54
The longest pause	752709.005	330978.25	204165.91	427540.005	240226.75
Location of the longest pause	2	3.5	4.90	1	8.5
Pause variance	2.70E+08	1.64E+08	5.66 E+07	7.55E+07	1.17E+08
Total pause time	3607.6	3855.74	2664.58	3758.58	2416.21
Total active writing time	1016.83	1026.19	1105.97	1512.17	801.62
Total process time	4624.43	4881.93	3770.55	5270.75	3217.83
Total words in main document	365	330.25	428.45	649	293.5
Mean word length	4.72	5.01	4.87	4.726	4.98
Mean typed in p-bursts	17.69	9.09	13.08	14.843	12.12
Total typed per minute	36.04	25.16	41.28	43.75	34.78
Total keystrokes	2859	2068.5	2596.818182	3846	1879.25

Table 4 Statistical significance of process indicators difference in pairs of clusters

Compared clusters	G M within-word pauses	G M before-word pauses	G M before-sentence pauses	The longest pause	Location of the longest pause	Pause variance	Total typed per minute
1 & 2	0.22	1.03E−14	1.39E−17	0.01	1.60E−29	1.17E−05	2.63E−25
1 & 3	0.00017	0.025	1.87E−05	5.98E−20	0.0004	5.97E−20	1.87E−05
1 & 4	1.50E−11	1.50E−11	1.50E−11	5.98E−20	6.80E−12	1.41E−11	1.50E−11
1 & 5	0.0006	0.0003	0.03	1.87E−32	3.59E−06	1.37E−17	1.55E−59
2 & 3	0.05	8.69E−41	2.62E−36	1.39E−17	4.25E−19	1.95E−54	0.0004
2 & 4	0.30	0.003	1.39E−17	1.39E−17	4.25E−19	1.36E−17	1.39E−17
2 & 5	0.09	9.07E−29	0.49	2.58E−07	7.75E−33	2.57E−07	3.57E−41
3 & 4	0.29	7.49E−14	3.93E−09	1.50E−11	8.42E−15	3.93E−09	1.50E−11
3 & 5	0.001	3.56E−06	1.80E−09	1.39E−17	8.04E−06	1.21E−07	1.39E−17
4 & 5	1.43E−05	5.12E−10	1.18E−05	1.18E−05	1.57E−11	0.49	1.39E−17
Compared clusters	Total pause time	Total active writing time	Total process time	Total words in main document	Mean word length	Mean typed in p-bursts	Total keystrokes
1 & 2	1.80E−09	1.55E−59	2.63E−25	8.04E−06	4.95E−07	0.008	8.63E−60
1 & 3	5.98E−20	1.87E−05	5.98E−20	3.32E−20	0.003	3.02E−10	4.80E−14
1 & 4	5.98E−20	5.98E−20	5.98E−20	8.42E−15	0.39	0.003	4.47E−20
1 & 5	1.55E−59	0.0003	1.96E−54	3.09E−18	2.53E−05	5.43E−25	8.87E−50
2 & 3	1.39E−17	1.39E−17	1.39E−17	7.44E−55	5.27E−06	5.01E−14	3.09E−18
2 & 4	1.39E−17	1.39E−17	1.39E−17	3.09E−18	4.18E−07	8.47E−06	3.09E−18
2 & 5	3.57E−41	3.57E−41	3.57E−41	7.58E−12	0.18	8.48E−15	1.68E−17
3 & 4	1.50E−11	1.50E−11	1.50E−11	3.32E−20	0.003	6.98E−05	8.42E−15
3 & 5	0.49	0.49	1.18E−05	5.36E−60	0.0002	1.39E−17	3.09E−18
4 & 5	0.49	1.E−17	1.18E−05	3.09E−18	2.15E−05	9.48E−16	3.09E−18

Student profiles and writing quality

We examined if the clusters are different in the quality of writing. Considering the process indicators, we expected that strategic planners, rapid writers, and emerging planners rate higher in their written outcomes than other clusters.

The results from the Mann-Whitney *U* test (Table 5) illustrate that the clusters differed on their means and standard deviations for the human ratings of their written outcome. Figure 7 depicts boxplots for the writing quality of each cluster. We further explored if these differences are statistically significant. As illustrated in Table 6, the difference in the writing quality between each pair of clusters is statistically significant.

Discussion

In this study, we aimed to identify clusters of students based on the keystroke data and the relationship between the clusters and writing quality. Specifically, we wanted to know (RQ1) if distinct clusters are revealed based on process indicators and (RQ2) whether the clusters of KL are related to writing quality.

Identifying student clusters

Based on previous studies, the heterogeneous nature of students writing behavior does not predict writing outcomes beyond a specific writing session or task (Conijn et al., 2021). Clustering techniques have only recently been used to account for this limitation and to inform process-oriented AWE tools design more efficiently compared with other approaches employed so far.

Previous research on clustering and writing quality prediction used some KL features for analysis. We select those features together with some new ones at the feature selection stage. Clustering analysis identifies fourteen of these KL features valuable for discovering student profiles. The analyses discover five clusters (Emerging planners, Average writers, Rapid writers, Strategic planners, and Low-performing writers (Research Question 1) and validate them with the final score. The number of identified clusters is in contrast with previous studies. For example, Zhang et al. (2017) found four student profiles using the mean essay score, total time spent on task, and the total number of words. The value of KL features they used for clustering is partially consistent with our findings which shows a high value for time on task and number of words in clustering. In addition, we validate the identified clusters with their writing quality while Zhang et al. (2017) used essay score as a feature for clustering. The value of this feature is not supported here. Shen and Chen (2021) also found two profiles based on KL features related to students' pausing behavior. They found that that less-skilled clusters paused more at word boundaries, but more-skilled clusters paused at sentence

Table 5 Descriptive statistics for clusters' writing quality

Cluster	Student profile	Mean	*SD*	Count	Writing quality
1	Emerging planners	21.90	0.22	30	Upper-intermediate
2	Average writers	19.82	2.27	120	Intermediate
3	Rapid writers	25.62	1.93	330	High
4	Strategic planners	29.90	0.12	30	Very high
5	Low-performing writers	17.61	0.61	120	Low

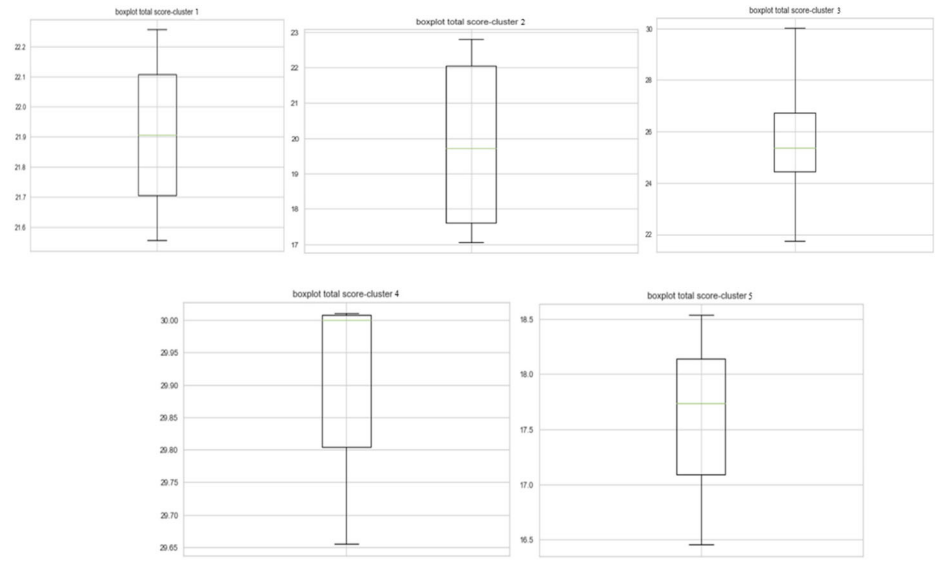

Fig. 7 The boxplot visualization for the writing quality of each cluster

boundaries. None of these were found here. In contrast, clustering analysis shows that high-achieving clusters (clusters 4, 3, and 1) have shorter pauses at sentence boundaries.

The differences in findings can be discussed in terms of the task type employed. Each specific task requires specific writing behaviors, and accordingly, different keystroke data is produced. The existing literature also indicated that keystroke data is sensitive to even minor differences in writing tasks, such as different prompts or source usage (Conijn et al., 2019; Guo et al., 2018; Sinharay et al., 2019). The differences are more explicit here as we used a different genre, i.e., argumentative writing tasks, while previous studies used source-based writing, online discussion, or academic summary tasks. Therefore, keystroke data and writing quality associations might also differ across tasks.

Table 6 Mean difference between clusters and their statistical significance

Cluster A	Comparison cluster	Mean difference of writing score	*p*-value
1	2	2.08	0.0002
1	3	3.71	5.28E−18
1	4	7.99	1.35E−11
1	5	4.29	1.39E−17
2	3	5.80	1.58E−54
2	4	10.08	1.38E−17
2	5	2.20	2.88E−10
3	4	4.28	8.21E−17
3	5	8.006	1.55E−59
4	5	11.42	1.38E−17

Student clusters and writing quality

After clustering analyses, correlational analysis was conducted to model the relationship between the identified clusters and the writing quality. The results indicated that varied pausing behavior and cognitive effort in clusters resulted in different writing qualities.

Our findings partially support previous studies, which found that some KL features predict the quality of the final product. For example, Sinharay et al. (2019) found that time on task, typing speed, number of bursts, and burst length correlated well with the quality of the final product. We found support for their finding as clusters 3 and 4, with the highest writing quality (Rapid writers and Strategic planners), show a long time on task and high typing speed. Another feature that showed a high effect size was the number of keystrokes (Choi & Deane, 2021). Our findings also support a high effect for the total number of keystrokes for Rapid writers and Strategic planners. In the following sections, we discuss the emerged clusters on their pausing behavior, cognitive effort in more detail.

Pausing behavior

Following previous studies researching writing quality prediction, we included pausing behavior keystroke features in clustering analysis. Out of the five identified clusters, three (Strategic planners, Emerging planners, Rapid writers) explained quantitative differences in the overall writers' pausing behavior. They differed on the geometric mean of before-sentence pauses, the longest pause, location of the longest pause, pause variance, and total pause time. specifically, the contrast in pausing behavior among clusters 1 (Emerging planners), 3 (Rapid writers), and 4 (Strategic planners) is noticeable. These clusters are statistically different in all three aspects of writing behavior we explored in this work (pausing behavior, fluency, time on task, and final product). However, the results indicate that the most significant difference among clusters 1, 3, and 4 lies in their pausing behavior.

Cluster 4 (who achieved the highest essay score) had their longest pause right at the beginning of the writing process (first interval), then continued with a signicant variance (7.55E+07) in the length of their pauses. In other words, their writing process was characterized by pauses of varied length.

Contrarily, cluster 3 started their writing process rapidly, i.e., without a pause, and left their longest pause, which is much shorter than cluster 1 and cluster 4 pauses, for the middle of their writing process. Their writing processes depict varied length pauses. However, the lengths of their pauses are not as varied as in cluster 4. Additionally, they finished the task much faster and with a much shorter total pause time than clusters 1 and 4.

Finally, cluster 1, Emerging planners, displayed some pausing behaviors prevalent in Strategic planners' profiles (cluster 4). Like Strategic planners, Emerging planners had a long pause at an initial stage (second interval) of their writing. However, they continued the whole writing process with almost same-lengthed pauses (variance = 2.704550e+08). These three clusters got the highest essay scores. One hypothesis is that these high-achieving clusters employ specific pausing strategies to deal with a complex task like argumentation. It seems that using initial long pause and pause-length variance, high-achievers managed to handle this complex task. To illustrate, argumentative

writing tasks need choosing a position, understanding the audience, researching the subject, and identifying the most convincing evidence for the opposing view. To effectively deal with this complexity, there needs to be a plan. Initial planning can help deal with these tasks, but constant planning during the writing process can be more helpful. This finding is also in line with Hayes' (2012) opportunistic view of planning: "Some writers tended to do all their planning before they began to write, and others interleaved planning with writing" (Hayes, 2012, p. 373). A helpful follow-up study would assess this assumption and examine if varied-length pauses throughout writing can explain the increase in writing achievement. The Average writers and Low-performing achievers did not show any specific pausing strategies. Their longest pause was less than average and happened at later stages of their writing process. In addition, they displayed a very low pause-length variance all over their writing process that puts across the idea that there were no strategies to deal with the lack of initial planning. They also achieved the lowest scores on their final product. Of note, total pause time for Average writers is similar to that of Strategic planners. However, considering pausing behaviors indicators altogether, we noticed that Strategic planners displayed different pausing patterns. Contrarily, Low-performing writers exhibited markedly lower total pause time compared with other profiles. This finding is in contrast with the findings of previous studies that related frequent pauses to low performance on writing (Shen & Chen, 2021). In short, the pausing pattern might be a more sensitive marker of successful performance than mere overall process indicators.

Cognitive effort

We can explain the differences between clusters by the variability in cognitive effort required for an argumentative writing task. Our findings indicate that the identified clusters are different in indicators of fluency, time-on-task, and product. Based on the Cognitive Load Theory (Brünken et al., 2010), we can explain the differences by different degrees of students' access to the relevant cognitive schemata. For example, specific schemata seem to be at hand for Strategic planners. Therefore, they can show optimal fluency levels and time on task to deal with a complex task (here: argumentation). On the other hand, rapid writers might have a set of general cognitive schemata that they can adjust to their needs during the argumentative writing task. This point gives us the notion that calling on the schemata entails some time to function reflected in their lower levels of fluency.

In the case of emerging planners, we observed that, on average, they had around average time-on-task, final product, and fluency indicators. These students were successful at completing the task and were not cognitively overloaded. However, the complexity of argumentation seems to surpass their cognitive capacity. Consequently, they did not score very high on the quality of their argumentation.

The limited time-on-task and production for low-performing writers seem to align with the notion of Ego depletion in Self-regulation Theory (Hagger et al., 2010). According to this theory, self-regulation is a restricted resource. If this source is drained,

it will lead to ego depletion. Consequently, the student will experience a decrease in cognitive performance. In line with Zhang et al.'s (2017) finding, this low performance seems not to result from cognitive overload but rather from low motivation.

In sum, our findings show that more proficient writers tend to handle their writing process more proficiently, produce text more fluently, and show more task engagement. On the contrary, low-performing students show less engagement with the task, less fluency, unsystematic pausing behavior, and produce less efficient essays.

The findings of the present work can be employed for diagnostic feedback and addressing students' difficulties. However, before direct use of AWE tools, further studies are required.

Limitations

There are some limitations in the present work that needs to be considered. First, the participants were not of diverse language background and age range. Second, we only used one type of task (argumentation). We might need different task types to discover general patterns used by students beyond task type. Third, longitudinal studies can detect students' interaction patterns with writing tasks and how they are affected by external variables or, in turn, affect writing quality.

Implications for AWE and future research

Our findings are a starting point for designing process-oriented AWE tools based on KL data. Future studies can build upon the current findings and examine bigger and more diverse samples to find intra-individual patterns of interaction with writing tasks and how students might move between clusters in the long run. Besides, diverse educational groups (e.g., elementary school, graduate, or post-graduate students) can be studied to find potentially different patterns of interactions and optimal points for providing feedback. Given the increasing importance of writing in peoples' lives and the need for AWE tools for self-regulation in writing, we invite researchers to plan and conduct such studies in the future.

Conclusions

The present work identified five student profiles based on their process behavior during an argumentative writing task. Three of these profiles exhibited quantitative differences in their pausing behavior. Information on how students manage their pausing behavior and their qualitative difference in their cognitive effort can improve AWE tools, which offer different types of feedback based on the students' background. We admit that much research is needed before using this information in AWE tools. We also believe it is essential to focus more on quantitative and qualitative differences among different groups of students, specifically in general thinking skills, such as planning, that can be used in various writing genres.

Appendix

Argumentative Essay Rubric				
Categories & Criteria - each worth eleven points	Level 4	Level 3	Level 2	Level 1
Thesis	Written with a clear and outstanding thesis.	Written with a clear thesis.	Written with a confusing or misleading thesis.	Missing a thesis.
Transitions	The writer uses effective words throughout the article to make transitions between ideas.	Some of the transitions are weaker, detracting from the writing and organization.	Some sections are too isolated – not linked by transitions.	Writing lacks transitions, leading to a disjointed and confusing reading.
Use and Evaluation of Sources	Utilizes three sources (at least one print source). Sources are reputable.	Utilizes three sources (all are online). Sources aren't quite reputable and/or aren't quite fitting for the topic.	Utilizes two sources. Sources are basically irrelevant and aren't acceptable.	Doesn't use sources.
Audience, Tone, and Rhetorical Appeals	All the reasons are written to convince the appropriate audience. Purpose has been achieved. Tone is consistent and convincing.	Some of the writing would concern or appeal to the intended audience. At times, the focus wanders from the intended audience. Tone is inconsistent at times.	Very little of the article contains reasons that would concern or appeal to the intended audience.	None of the article contains arguments and/or reasons that address the intended audience.
Organization	Arguments are organized logically and coherently.	At times, the argument is not logically organized and/or the evidence doesn't support the claims.	Very little of the essay is well organized. Claims are not supported by evidence.	Arguments are illogically organized and incoherent.
Claims, Warrants, and Support	Writing addresses assumptions, makes at least three smaller claims re: the major argument, and provides support for every claim.	One or more claim is lacking support and the writer doesn't clearly address assumptions.	Writing doesn't address assumptions, makes only two or fewer claims re: the major argument and/or doesn't provide support.	Writing only makes one major claim and doesn't address any assumptions.
Paraphrase, Direct Quotation, and Summary	Writing contains a balanced and successful mix of paraphrase, direct quote, and summary.	Writing only contains two of the three and/or paraphrase and/or summary are done incorrectly.	Writing only contains one of the three or they are done incorrectly.	Writing lacks all three.
In-text Citations and Works Cited Page	Using MLA format, correctly cites all sources used on the works cited page. Incorporates quotes correctly in the essay.	Using MLA format, correctly cites all sources on the works cited page with only a few minor errors. Incorporates quotes correctly in the essay with only a few minor errors.	Incorrectly cites sources on the works cited page and doesn't correctly incorporate quotes in the body of the essay.	Doesn't cite sources used or doesn't use sources.
Mechanics	Uses all correct grammar and spelling throughout. Sentence variety and word choices are outstanding. Doesn't use "I" or first person POV, except in sections of personal narrative.	Uses mostly correct grammar and spelling. Some attempt at variety in words choice/sentence variety. Doesn't use "I" or first person POV, except in sections of personal narrative.	Several grammar and spelling mistakes. Words choices are simple; sentences lack variety. Uses "I" or first person POV sparingly.	Many grammar and spelling mistakes. Word choices are weak and sentence variety is nonexistent. Is written from first person POV.
Style and Syntax	Uses varied sentence length and structure, and has a mature, college-level style.	Attempts to use varied sentence length and structure, and is almost at a mature, college-level style.	Overuses short or long sentences, has similar sentence structure throughout, and has high-school level style.	Lacks appropriate syntax and any attempt at style.

Abbreviations

AWE: Automated writing evaluation; KL: Keystroke logs; ML: Machine learning; MUST: Mazandaran University of Science and Technology

Acknowledgements

Not applicable.

Authors' contributions

Mobina Talebinamvar mainly contributed in coding, programming, and the analysis process. Forooq Zarrabi contributed in study design, data collection, algorithm design, analysis, and interpretation of the results. She was also the major contributor in writing the manuscript. Both authors read, revised, and approved the final manuscript.

Availability of data and materials

The datasets used during the current study are available from the corresponding author upon request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Engineering college, Alzahra University, Vanak Village Street, Tehran 19938 93973, Iran. [2]Department of Languages, Mazandaran University of Science and Technology, Sheikh Tabarsi Street, Sardaran 12, Babol, Iran.

Received: 23 August 2021 Accepted: 27 November 2021
Published online: 07 February 2022

References

Allen, L. K., Jacovina, M. E., Dascalu, M., Roscoe, R. D., Kent, K. M., Likens, A. D., & McNamara, D. S. (2016). {ENTER} ing the time series {SPACE}: Uncovering the writing process through keystroke analyses. *International Educational Data Mining Society*.

Al-Saadi, Z. T., & Galbraith, D. (2020). Does the revision process differ across the language of writing (L1 vs. FL), FL language proficiency, and gender? An empirical study using keystroke logging data. *Writing and Pedagogy*.

Baaijen, V. M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*, *36*(3), 199–223. https://doi.org/10.1080/07370008.2018.1456431.

Barkaoui, K. (2019). What can L2 writers' pausing behavior tell us about their L2 writing processes. *Studies in Second Language Acquisition*, *41*(3), 529–554. https://doi.org/10.1017/S027226311900010X.

Bowen, N., & Van Waes, L. (2020). Exploring revisions in academic text: Closing the gap between process and product approaches in digital writing. *Written Communication*, *37*(3), 322–364. https://doi.org/10.1177/0741088320916508.

Brooks, M., Basu, S., Jacobs, C., & Vanderwende, L. (2014). Divide and correct: Using clusters to grade short answers at scale. In *Proceedings of the first ACM conference on Learning@ scale conference*, (pp. 89–98).

Brünken, R. E., Plass, J. L., & Moreno, R. E. (2010). Current issues and open questions in cognitive load research. In J. L. Plass, R. Moreno, & R. Brunken (Eds.), Cognitive load theory (pp.253e272). New York: Cambridge University Press.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354.
Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in higher education*, *36*(4), 395–407. https://doi.org/10.1080/03075071003642449.
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.
Chou, C. Y., & Zou, N. B. (2020). An analysis of internal and external feedback in self-regulated learning activities mediated by self-regulated learning tools and open learner models. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–27. https://doi.org/10.1186/s41239-020-00233-y.
Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H. H. (2019). Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition*, *41*(3), 583–604. https://doi.org/10.1017/S027226311900007X.
Conijn, R., Cook, C., van Zaanen, M., & Van Waes, L. (2021). Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, 1–32.
Conijn, R., Martinez-Maldonado, R., Knight, S., Buckingham Shum, S., Van Waes, L., & Van Zaanen, M. (2020). How to provide automated feedback on the writing process? A participatory approach to design writing analytics tools. *Computer Assisted Language Learning*, 1–31.
Conijn, R., van Zaanen, M., & Van Waes, L. (2019). Don't wait until it is too late: The effect of timing of automated feedback on revision in ESL writing. In *European Conference on Technology Enhanced Learning*, (pp. 577–581). Cham: Springer.
Cunningham, K. J. (2019a). Student perceptions and use of technology-mediated text and screencast feedback in ESL writing. *Computers and Composition*, *52*, 222–241. https://doi.org/10.1016/j.compcom.2019.02.003.
Cunningham, K. J. (2019b). How language choices in feedback change with technology: Engagement in text and screencast feedback on ESL writing. *Computers & Education*, *135*, 91–99. https://doi.org/10.1016/j.compedu.2019.03.002.
Cunningham, K. J., & Link, S. (2021) Video and text feedback on ESL writing: Understanding attitude and negotiating relationships. *Journal of Second Language Writing, 52*, 100797.
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.
Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2019). Writing process differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics*, *44*(5), 571–596. https://doi.org/10.3102/1076998619856590.
Hagger, M. S., Wood, C. W., Stiff, C., & Chatzisarantis, N. L. (2010). Self-regulation and self-control in exercise: The strength-energy model. *International Review of Sport and Exercise Psychology*, *3*(1), 62–86. https://doi.org/10.1080/17509840903322815.
Hayes, J. R. (2012). Modeling and remodeling writing. *Written communication*, *29*(3), 369–388. https://doi.org/10.1177/0741088312451260.
Henderson, M., Ryan, T., & Phillips, M. (2019). The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education.*, *44*(8), 1237–1252. https://doi.org/10.1080/02602938.2019.1599815.
Hung, J. L., Wang, M. C., Wang, S., Abdelrasoul, M., Li, Y., & He, W. (2015). Identifying at-risk students for early interventions—A time-series clustering approach. *IEEE Transactions on Emerging Topics in Computing*, *5*(1), 45–55. https://doi.org/10.1109/TETC.2015.2504239.
Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011.
Kochmar, E., Do Vu, D., Belfer, R., Gupta, V., Serban, I. V., & Pineau, J. (2020). Automated personalized feedback improves learning gains in an intelligent tutoring system. In *International Conference on Artificial Intelligence in Education*, (pp. 140–146). Cham: Springer.
Lam, D. M. (2021). Feedback as a learning-oriented assessment practice: Principles, opportunities, and challenges. In *Learning-Oriented Language Assessment*, (pp. 85–106). Routledge.
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.
Likens, A. D., Allen, L. K., & McNamara, D. S. (2017). Keystroke dynamics predict essay quality. In *CogSci*.
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.
Mojarad, S., Essa, A., Mojarad, S., & Baker, R. S. (2018). Data-driven learner profiling based on clustering student behaviors: Learning consistency, pace and effort. In *International Conference on Intelligent Tutoring Systems*, (pp. 130–139). Cham: Springer.
Otnes, H., & Solheim, R. (2019). Acts of responding. Teachers' written comments and students' text revisions. *Assessment in Education: Principles, Policy & Practice*, *26*(6), 700–720.
Ranalli, J. (2019). Automated written corrective feedback for supporting students and instructors across curriculum. In CELT Teaching Technology Conference (Vol. 2019, No. 1). Iowa State University Digital Press..
Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, *52*, 100816. https://doi.org/10.1016/j.jslw.2021.100816.
Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed-methods study. *Studies in Second Language Acquisition*, *41*(3), 605–631. https://doi.org/10.1017/S027226311900024X.
Saricaoglu, A. (2018). The impact of automated feedback on L2 learners' written causal explanations. *ReCALL, 1 of 15*, *31*(2), 189–203. https://doi.org/10.1017/S095834401800006X.
Sarid, M., Peled, Y., & Vaknin-Nusbaum, V. (2021). The relationship between second language college students' perceptions of online feedback on draft-writing and academic procrastination. *Reading and Writing*, *34*(5), 1247–1271. https://doi.org/10.1007/s11145-020-10111-8.
Shen, C., & Chen, N. (2021). Profiling the pausing behaviour of EFL learners in real-time computer-aided writing: a multi-method case study. *Asian-Pacific Journal of Second and Foreign Language Education*, *6*(1), 15. https://doi.org/10.1186/s40862-021-00118-1.

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.
Steen-Utheim, A., & Hopfenbeck, T. N. (2019). To do or not to do with feedback. A study of undergraduate students' engagement and use of feedback within a portfolio assessment design. *Assessment & Evaluation in Higher Education*, *44*(1), 80–96. https://doi.org/10.1080/02602938.2018.1476669.
Tseng, S.-S., & Yeh, H.-C. (2019). The impact of video and written feedback on student preferences of English-speaking practice. *Language Learning & Technology*, *23*(2), 145–158.
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), 109–139. https://doi.org/10.17239/jowr-2020.12.01.05.
Wallot, S., & Grabowski, J. (2019). A tutorial introduction to recurrence quantification analysis (RQA) for keystroke logging data. *Observing Writing*, 163–189.
Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. *Handbook of formative assessment*, 18–40.
Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education*, *168*, 104208. https://doi.org/10.1016/j.compedu.2021.104208.
Winstone, N. E., & Boud, D. (2020). The need to disentangle assessment and feedback in higher education. *Studies in Higher Education*, 1–12.
Yamashita, T., & Ranalli, J. (2020). Corrective feedback in collaborative writing: Do variations in contributed revisions predict variations in learning. In *2020 Conference of the American Association for Applied Linguistics (AAAL)*.
Zarrabi, F., & Bozorgian, H. (2020). EFL students' cognitive performance during argumentative essay writing: A log-file data analysis. *Computers and Composition*, *55*, 102546. https://doi.org/10.1016/j.compcom.2020.102546.
Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and psychological measurement*, *76*(2), 280–303. https://doi.org/10.1177/0013164415590022.
Zhang, M., & Deane, P. (2015). Process features in writing: Internal structure and incremental value over product features. *ETS Research Report Series*, *2015*(2), 1–12. https://doi.org/10.1002/ets2.12075.
Zhang, M., Zhu, M., Deane, P., & Guo, H. (2017). Identifying and comparing writing process patterns using keystroke logs. In *The Annual Meeting of the Psychometric Society*, (pp. 367–381). Cham: Springer.
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.
Zheng, L., Zhong, L., & Niu, J. (2021). Effects of personalised feedback approach on knowledge building, emotions, co-regulated behavioural patterns and cognitive load in online collaborative learning. *Assessment & Evaluation in Higher Education*, 1–17.

Publisher's Note