# {ENTER}ing the Time Series {SPACE}: Uncovering the Writing Process through Keystroke Analyses

Laura K. Allen[1], Matthew E. Jacovina[1], Mihai Dascalu[2], Rod D. Roscoe[1], Kevin M. Kent[1], Aaron D. Likens[1], & Danielle S. McNamara[1]
[1]Arizona State University, Tempe, USA
[2]University Politechnica of Bucharest, Bucharest, Romania
{LauraKAllen, Matthew.Jacovina}@asu.edu, {Mihai.Dascalu}@cs.pub.ro, {Rod.Roscoe, Kkent4, Aaron.Likens, dsmcnama}@asu.edu

## ABSTRACT
This study investigates how and whether information about students' writing can be recovered from basic behavioral data extracted during their sessions in an intelligent tutoring system for writing. We calculate basic and time-sensitive keystroke indices based on log files of keys pressed during students' writing sessions. A corpus of prompt-based essays was collected from 126 undergraduates along with keystrokes logged during the session. Holistic scores and linguistic properties of these essays were then automatically calculated using natural language processing tools. Results indicated that keystroke indices accounted for 76% of the variance in essay quality and up to 38% of the variance in the linguistic characteristics. Overall, these results suggest that keystroke analyses can help to recover crucial information about writing, which may ultimately help to improve student models in computer-based learning environments.

## Keywords
Intelligent Tutoring Systems; Writing; Natural Language Processing; Feedback; Keystrokes; Temporality

## 1. INTRODUCTION
Effective written communication is a complex socio-cognitive skill that is important for success in academic and professional settings [1-2]. The writing process relies on both lower- and higher-level knowledge and skills, ranging from knowledge of the language and domain to strategies necessary for generating inferences and flexibly adapting to different task demands [1; 3-5] Not surprisingly, then, the development of strong writing skills is extremely difficult and students consistently underachieve on national and international assessments of writing [6-8].

The remediation of these writing deficits is a similarly challenging task. The development of writing proficiency demands that students have access to high-quality instruction that is attuned to their particular needs. Research on writing instruction finds that students attain the greatest benefits when they are provided strategy instruction, practice, and feedback [9-10]. In particular, *deliberate practice* is crucial for the development of writing skills [11] and has been shown to help students regulate the planning, drafting, and reviewing stages of writing [10]. This type of meaningful and mindful practice inherently relies upon individualized formative feedback—feedback that reveals and explains actionable steps that students must take to improve. However, in large classrooms, detailed and targeted feedback on multiple essay drafts per student presents a daunting challenge for teachers.

Computer-based tools such as automated writing evaluation (AWE) systems have been developed to alleviate some of the pressures facing writing instructors [12]. At their core, AWE tools implement natural language processing (NLP) and machine learning techniques to accurately model the scores that expert human raters would assign based on the structure and content of students' essays [13-14]. Additionally, many AWE systems and intelligent tutoring systems (ITSs) incorporate instructional elements such as lessons and practice games [15-16]. These modern systems extend beyond the assessment of essay quality to provide students with personalized feedback and recommendations for improvement.

Although a wealth of research has been conducted to validate the *accuracy* of AWE scores, much less attention has been paid to the pedagogical and rhetorical elements of these systems. Specifically, critics often cite the lack of sensitivity to different audiences, rhetorical moves, and writing processes as serious areas of concern, which can lead to impersonal and ineffective instruction and feedback [17;18]. These critiques are valid and point to much needed future research. Accordingly, researchers and developers have begun to re-focus their efforts away from establishing the accuracy of scoring models and towards the improvement of the personalized and nuanced aspects of the feedback and instruction.

To better detect and respond to differences among students' writing processes and behaviors, we may need to embed assessments that are based on more than their written products and essay scores. These measures can be either visible or hidden from users (i.e., "stealth assessments") [19], and can inform specific instruction and feedback that is tailored to students' individual habits. In the context of computer-based learning environments, these assessments can be informed by a wealth of information that is easily logged within the system. Snow and colleagues (2014) [20], for example, developed stealth assessments of self-regulation within a reading comprehension tutoring system. They found that the predictability of students' choices in the system was

indicative of their self-regulation skill and influenced their performance on the learning task. Overall, such assessments may offer a viable solution to the writing process assessment problem. Both simple measures (e.g., typing speed) and complex measures (e.g., trajectories of mouse movements) might allow us to model the writing processes and characteristics of student users.

In this paper, we examine the efficacy of behavioral measures that are accessible (but rarely collected or analyzed) in writing training systems to detect information about students' performance on their essays. In particular, we examine whether basic and time-sensitive keystroke indices can be used to model the scores and linguistic features of students' essays. Our ultimate goal is to use these models to provide more individualized tutoring and feedback to students.

## 1.1 Keystroke Analyses for Writing

Keystroke data presents a potentially valuable approach for modeling students' writing behaviors [e.g., 21]. Although researchers have made significant strides in leveraging the linguistic features of texts to understand writing quality, there has been substantially less research on students' online or real-time writing processes. Due to challenges of data collection, prior writing research has focused primarily on students' finished writing products and not their moment-by-moment writing processes. Recently, however, keystroke logging tools (i.e., software that records the keys individuals press while typing) have been applied to the study of writing [22]. These tools offer a viable way to study students' actions as they compose and edit their essays. One such tool, InputLog, has been developed to interface with NLP tools, which enables analyses that synthesize both keystroke and linguistic data.

Illustrative examples of the value of keystroke analyses stem from work on affect detection during writing [21; 23]. Writers' affective states during writing—ranging from boredom and frustration to excitement and engagement—can have a significant impact on the writing experience and eventual products. However, these qualities may not be detectable from written products alone. How might keystroke patterns vary when writers are in a fluid, engaged "flow" state as compared to a frustrated struggle to generate ideas?

In recent work, Bixler and D'Mello (2013) [21] have begun to explore such questions. They collected individual difference measures and keystroke data from student writers to detect online affective states during writing (i.e., self-reported affective states in 15-second intervals). Their results indicated that a combination of behavioral (keystroke) measures and student-level indices was able to detect boredom, engagement, and neutral states between 11% and 38% above baseline. Similarly, Allen et al. (in press) [23] combined individual difference, linguistic, and keystroke indices to predict engagement and boredom across writing sessions. Their results suggested that these three categories of indices were successful in modeling students' affective states during writing. Indices related to academic ability, text properties, and keystroke logs were able classify high and low engagement and boredom in writing sessions with 77% accuracy.

In sum, keystroke analyses hold the potential to reveal crucial data on students' online writing experiences and processes that are normally invisible in product-based analyses alone.

## 1.2 Writing Pal

A long-term goal of our research is to improve personalized, adaptive learning and feedback within the Writing Pal (W-Pal) intelligent tutoring system [24]. W-Pal offers explicit strategy instruction, practice, and feedback for prompt-based persuasive essay writing for high school and early college students. Relative to other writing training systems (see [24] for a review), W-Pal is unique in its focus on explicit strategy instruction and its varied opportunities for practice (i.e., game-based strategy practice and essay writing practice). Strategy instruction is delivered via video presentations on canonical writing processes: prewriting, drafting, and revising. These videos feature virtual pedagogical agents who explain and demonstrate a variety of principles and strategies (see Figure 1 for a screenshot of the Freewriting Module). These lessons include: Freewriting and Planning (prewriting); Introduction Building, Body Building, and Conclusion Building (drafting); and Paraphrasing, Cohesion Building, and Revising (revising). After completing lessons, students unlock a suite of strategy practice mini-games. In these games, students reinforce their strategy knowledge through both generative and identification tasks. Game-based practice allows students to work on specific components of the writing process and strategies prior to applying them in a complete essay composition.

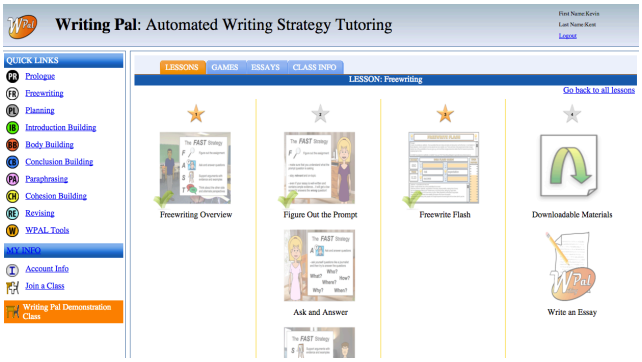


**Figure 1. Screenshot of the the Freewriting module**

### *1.2.1 W-Pal Essay Practice and Feedback*

W-Pal also gives students the opportunity to practice writing persuasive essays and receive summative and formative feedback. Writing takes place in a word-processing interface where students can view the prompt, a "scratch-pad" for brainstorming and outlining, and the writing space. Once the essays are submitted, a combination of formative and summative feedback is provided. Like other AWEs, W-Pal employs NLP tools to extract linguistic data from essays, and implements a series of algorithms to assess quality and guide feedback delivery. In analyzing the text, the system considers characteristics across a variety of linguistic indices.

Summative feedback (see Figure 2) includes a holistic score on a 1-6 scale, with descriptors representing each level (i.e. "Great"). Formative feedback (see Figure 2) is given both at the essay-level (i.e. length, relevance, structure) and section-level (i.e. suggestions to improve an introduction). This formative feedback is designed to be specific, actionable, and aligned to strategies taught in the lessons. For example, students who submit essays with weak conclusions may receive feedback about summarizing key arguments from the body paragraphs in the conclusion. After viewing the feedback, students can revise their essays. In the

revision phase, essay feedback is displayed adjacent to the writing space, facilitating uptake of the recommendations.

Previous research evaluating the efficacy of the W-Pal system has found that this training results in improved essay scores, increased strategy knowledge, and improved revising strategies [15; 25-26].

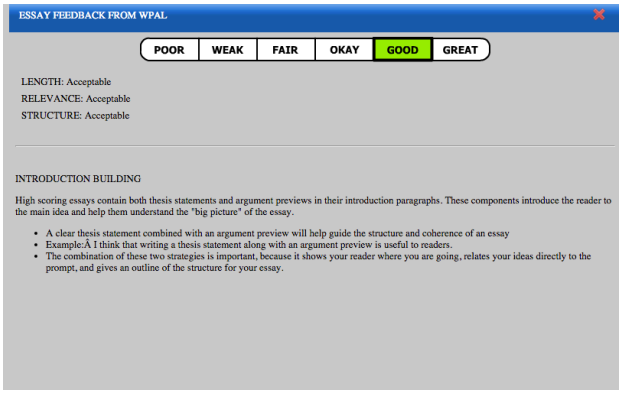


**Figure 2. Screenshot of the feedback window**

## 1.3 CURRENT STUDY

The current study investigates how and whether information about students' writing behaviors within W-Pal can be recovered from basic behavioral data extracted from keystroke analyses. To this end, we calculate a number of indices based on the keystrokes pressed by student writers with the intent of modeling the quality and linguistic features of their essays. An overarching aim of this research is to develop online, stealth assessments of students' writing processes that can inform new student models and system adaptivity. An increase in the sensitivity of W-Pal to students' writing processes is expected to improve its ability to offer more nuanced and personalized feedback and recommendations.

We collected timed, persuasive essays written by undergraduate students and scored using the W-Pal algorithm [27]. Linguistic properties of the essays were assessed via Coh-Metrix [28] and WAT [29], which are automated NLP tools that calculates text information related to lexical, syntactic, cohesive, and rhetorical properties. In addition, we logged keystrokes during students' writing session and calculated measures related to the general and temporal properties of these keystroke logs.

We hypothesized that these basic and time-sensitive keystroke indices would provide meaningful information about the writing processes enacted by students, which would subsequently relate to the quality and characteristics of their essays.

# 2. METHODS

## 2.1 Participants

We recruited 131 undergraduate participants from a university in the United States, who received course credit. Students reported a mean age of 19.8 years, with 44.3% identifying as female, 64.1% Caucasian, 14.5% Asian, 7.6% African American, 7.6% Hispanic, and 6.1% as "Other." Data for five students were lost due to computer error; thus, the final corpus comprised 126 essays.

## 2.2 Data Collection Procedure

Participants wrote a timed (25-minute), prompt-based, persuasive essay. Essay prompts resembled typical SAT items, and students were not allowed to proceed until the full 25 minutes elapsed. Students typed their essays in the AWE component of W-Pal and all keystrokes were logged along with millisecond timestamps. Essays contained an average of 412.3 words (*SD* = 159.9, *min* = 47.0, *max* = 980.0).

## 2.3 Essay Scoring

Students' essays were automatically scored using a computational algorithm that assigns scores on a scale from 1 (lowest) to 6 (highest). This algorithm relied on linguistic features computed by Coh-Metrix, the Writing Assessment Tool (WAT), and Linguistic Inquiry and Word Count (LIWC). For more details on this algorithm, see [27].

## 2.4 Text Analyses

Linguistic properties of essays were assessed via two NLP tools: Coh-Metrix [28] and WAT [29]. These tools report hundreds of linguistic indices that relate to text structure, general readability, rhetorical patterns, lexical choices, and cohesion. For the current analyses, we selected four indices from Coh-Metrix and WAT that demonstrated theoretical ties to writing quality. We chose this limited number of indices to specifically examine whether and how the keystroke indices would map onto four key dimensions of the essays: lexical, syntactic, semantic, and cohesion.

**Word Frequency.** Coh-Metrix and WAT calculate multiple indices that describe the specific types of words used in texts. Word frequency measures, for instance, are used to assess how frequently certain words occur in the English language. Coh-Metrix reports indices of word frequency that are taken from the CELEX database. Additionally, Coh-Metrix reports the logarithm of word frequency for all words in a text. An index of log frequency is calculated because reading times are typically linearly related to the logarithm of word frequency rather than the raw word frequency [30]. For this reason, we chose to examine the log frequency of all words.

**Syntactic Complexity.** Additionally, Coh-Metrix and WAT contain a number of indices that describe the properties of the sentences in texts, such as the frequency of specific parts of speech and the complexity of their syntactic constructions. Sentence complexity is assessed by multiple indices. More complex syntax is typically associated with higher quality essays [28] and recent evidence suggests that working memory capacity is linked to the production of more complex syntax [31]. Here, we used the index mean number of words before the main verb as a proxy for sentence complexity.

**Semantic Diversity.** Semantic diversity refers to the number of unique concepts expressed in an essay. This measure is conceptually similar to measures of lexical diversity, but more strongly emphasizes the diversity of ideas rather than specific words. A semantic diversity score is calculated in WAT using Latent Semantic Analysis (LSA) [32] and is operationalized as the ratio of semantically independent concepts to the total number of word types in an essay.

**Global Semantic Cohesion.** Global semantic cohesion is also calculated in WAT using LSA. Here, we used the index LSA (start-to-end), which calculates the degree to which the introduction and conclusion of an essay contain semantically similar information. We chose this index (rather than examining the semantic similarity between all the paragraphs) because higher-quality essays typically share semantic content in the opening and closing paragraphs, but bring in outside information in the form of arguments and evidence in the body paragraphs.

## 3. KEYSTROKE ANALYSES

To investigate whether and how students' writing behaviors were related to the quality and linguistic properties of their essays, we computed a number of keystroke indices. In particular, we calculated both *basic keystroke indices* (i.e., indices that were aggregated across the entire essay), as well as *time-sensitive keystroke indices* (i.e., indices that accounted for the temporal nature of the keystroke data).

### 3.1 Basic Keystroke Indices

Basic keystroke indices aggregated the number of specific writing events (e.g., pauses and backspaces) that occurred across an entire writing session. These basic indices are deliberate replications of indices from previous studies because they have been successfully used to model students' affect during writing [21; 23]. Table 1 provides an overview of these indices.

**Table 1. Basic Keystroke Indices**

| Measure | Description |
|---|---|
| Verbosity | Number of keystrokes per essay |
| Backspaces | Number of backspaces per essay |
| Largest Latency | Largest time difference between keystrokes during essay writing |
| Smallest Latency | Smallest time difference between keystrokes during essay writing |
| Median Latency | Median of all the differences in time between keystrokes per essay (not including initial pause) |
| Initial Pause | Length of the first pause of an essay writing session |
| 0.5 Second Pauses | Number of pauses above .5 seconds and below 1 second |
| 1 Second Pauses | Number of pauses above 1 second and below 1.5 seconds |
| 1.5 Second Pauses | Number of pauses above 1.5 seconds and below 2 seconds |
| 2 Second Pauses | Number of pauses above 2 seconds and below 3 seconds |
| 3 Second Pauses | Number of pauses above 3 seconds |

### 3.2 Time-Sensitive Keystroke Indices

Despite the importance of basic keystroke indices, indices that aggregate behavioral patterns over the course of an entire essay session can miss out on important temporal variability. For instance, consider the time series depicted in Figure 3. This plot shows the number of keystrokes pressed by one student writer within each 30 second window of a writing session. The student clearly did not maintain stable behavioral patterns throughout the writing session; instead, she engaged in periods of high and low activity. Analyses that are restricted to basic indices necessarily ignore this variability. We hypothesize that investigations into the temporal structure of the keystrokes (i.e., the distributions of events in time) will provide meaningful information about students' writing processes beyond the basic aggregated measures.

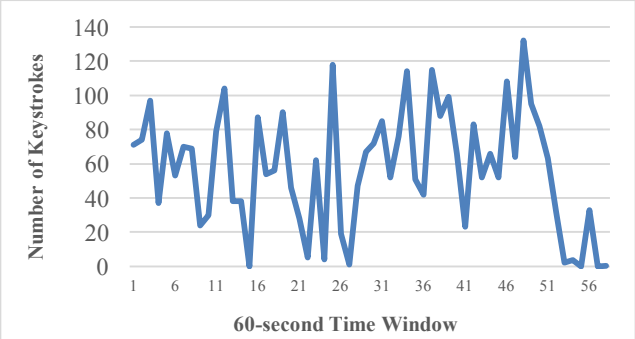


**Figure 3. Variability of keystroke patterns for a single student**

**Table 2. Time-Sensitive Keystroke Indices**

| | Description |
|---|---|
| StDev Events | Standard deviation of the number of events in each time window |
| Slope Degree | Slope of the linear regression applied on the time series |
| Entropy | Shannon's Entropy calculated for the number of events in the windows normalized by the total number of events for the overall time series. If a student only typed in a single window, the entropy would be 0. When maintaining a constant typing rate, entropy converges toward the maximum value of log(n). |
| Degree of Uniformity | Uniformity of the time series (Jensen-Shannon divergence method), which is a symmetric and bounded function of similarity that calculates the similarity between two distributions: a uniform probability distribution of 1/n (i.e., a constant typing rate) and the probability of key presses in a given window (i.e., the actual time series produced by the student). |
| Local Extremes | Number of time windows for which the direction of the evolution of keystroke events changes. This reflects inconsistency in writing rates across the windows. |
| Average Recurrence | Average recurrence of events across the time windows. This recurrence is expressed as the distances between time windows that contain at least one keystroke event. This measure is useful for identifying writing pauses. If each time window has at least one event, recurrence is 0, whereas if students take long pauses that occasionally result in time windows of 0 events, recurrence increases (if they write every two time windows, recurrence will be one). |
| StdDev Recurrence | Standard deviation of the recurrence across the time windows |

Note: All time-sensitive keystroke indices were calculated using 30- and 60-second time windows.

To this end, we calculated a number of new indices that we have classified as *time-sensitive keystroke indices*. These indices deliberately take the within-subject temporal distribution of keystroke events into account. The time series of keystrokes

generated during students' sessions were first separated into non-overlapping windows of 30 and 60 seconds to account for variability across different scales. These individual windows contained information about the number of keystroke events that occurred in each time window. The time-sensitive keystroke indices were then separately generated based on each of the two window intervals (see Table 2).

## 3.3 Statistical Analyses

Statistical analyses investigated whether basic and time-sensitive keystroke indices accounted for variability in student writing performance. Pearson correlations were first calculated between the holistic essay scores and the keystroke indices obtained from the writing sessions (see Tables 1 and 2). Indices that displayed a significant or marginally significant correlation with essay scores ($p < .10$) were retained in the analysis.

Normality of the indices was assessed with skew, kurtosis, and visual data inspections, and no indices were removed based on these inspections. Range transformations (0-1) were applied to ensure that the keystroke and linguistic indices were on the same scale. Multicollinearity was then assessed among the indices ($r > .90$). When two or more indices demonstrated multicollinearity, the index that correlated most strongly with essay scores was retained in the analysis.

A linear regression analysis[1] was conducted using M5-prime feature selection to assess which of the remaining keystroke indices were most predictive of essay scores. To avoid overfitting the model, we chose a ratio of 15 essays to 1 predictor, which allowed for a maximum of eight indices to be entered in to the model, given that there were 126 essays included in the analysis.

We first conducted the regression analysis on the entire corpus, and then validated the model using ten-fold cross-validation with shuffled sampling. In this cross validation analysis, the corpus was first split into 10 "folds" and each fold was individually removed from the corpus for each analysis and the remaining essays were used as the training set. We tested the accuracy of the linear regression model by examining its ability to model the omitted fold. The process was repeated until each fold was omitted once in the test set. This analysis therefore allowed us to test the model's accuracy on independent sets of data (i.e., data that are not in the training set). If the overall model and the model generated by the cross-validation analysis are similar, our confidence in model stability is increased.

Following this essay score analysis, similar follow-up analyses were conducted using the keystroke indices to predict the linguistic features of the essays. For these analyses, we followed the same procedure detailed above.

# 4. RESULTS

## 4.1 Keystrokes and Essay Quality

Pearson correlations were calculated between the basic and time-sensitive keystroke indices and students' holistic essay scores to examine the strength of the relationships among the variables. The correlation analysis revealed that there were 10 keystroke indices that demonstrated a significant relation with holistic essay scores and did not demonstrate multicollinearity with each other. To avoid overfitting the model, we only selected the eight indices that were most strongly correlated with essay scores. These eight indices are listed in Table 3.

**Table 3. Correlations between Essay Scores and Keystroke Indices**

| Keystroke Index | *r* | *p* |
|---|---|---|
| Verbosity | 0.819 | <.001 |
| Local Extremes (30s time window) | -0.476 | <.001 |
| Entropy (30s time window) | 0.472 | <.001 |
| Median Latency | -0.436 | <.001 |
| StdDev Events (30s time window) | 0.397 | <.001 |
| Largest Latency | -0.359 | <.001 |
| Backspaces | 0.308 | <.001 |
| StdDev Recurrence (30s time window) | -0.297 | = .001 |

A linear regression analysis was calculated with the eight keystroke indices as predictors of students' essay scores (score range: 1-6). This analysis yielded a significant model, $R^2 = .758$, $RMSE = 0.377$, $p < .001$, with three variables that combined to account for 76% of the variance in the essay scores: *Verbosity* [$\beta = 1.03$, $p < .001$], *Largest Latency* [$\beta = -.09$, $p < .001$], and *Backspaces* [$\beta = .39$, $p < .001$]. The follow-up ten-fold cross validation analysis produced a significant model with similar statistics, $R^2 = .737$, $RMSE = 0.386$.

An interesting question is whether additional indices provided useful information about the essay quality once Verbosity was removed from the analysis. That is, including the total number of key presses may suppress the important role of other writing behaviors. We conducted a second regression analysis that excluded Verbosity. This regression yielded a significant model, $r = .778$, $R^2 = .606$, $RMSE = 0.482$, $p < .001$. Six variables were significant or marginally significant predictors in the regression analysis and combined to account for 61% of the variance in students' essay scores: *StdDev Events* (30s) [$\beta = 0.529$, $p < .001$], *Entropy* (30s) [$\beta = 1.047$, $p < .001$], *StdDev Recurrence* (30s) [$\beta = -0.509$, $p < .001$], *Backspaces* [$\beta = 0.209$, $p < .01$], *Local Extremes* (30s) [$\beta = -0.176$, $p < .05$], and *Median Latency* [$\beta = -0.141$, $p = .096$]. As above, the cross validation model produced similar results, $R^2 = .588$, $RMSE = 0.534$.

In sum, these correlation and regression analyses indicate that better writers pressed more keys (both characters and backspace) over the course of their writing session. They also maintained a more consistent rate across the 30 second time windows (i.e., whether they typed or not within the individual time windows), as measured by Entropy, Local Extremes, and StdDev Recurrence indices, but exhibited greater variability in the *number of keystroke events* within the 30s time windows (StdDev Events). Additionally, these students' keystroke logs were characterized by shorter pause times as measured both by the Median and Largest Latency indices. Taken together, these findings demonstrate that *writing fluency*—the ease and consistency with which writers generate text—is a key indicator of proficiency (e.g., [33]). This work both confirms and extends prior research by investigating a

[1] We investigated the usefulness of a number of regression and neural net techniques in the current study. However, due to space limitations, these models are not reported. In the end, we report the the linear regression models because this approach yielded the strongest and most stable models.

feature of higher quality writing using process analyses rather than post-hoc linguistic analyses alone.

## 4.2 Keystrokes and Linguistic Features

Our second aim was to investigate whether keystroke indices were related to specific linguistic features of the essays. Pearson correlations were calculated between the keystroke indices and the four linguistic variables calculated by Coh-Metrix and WAT. These analyses were then followed by a regression analysis, and validated using ten-fold cross validation. The statistical information for these resulting models is provided below.

**Word Frequency.** The word frequency regression analysis yielded a significant model, $R^2$ = .185, *RMSE* = 0.179, $p < .001$. Three variables were significant or marginally significant predictors: *2 Second Pauses* [β = -0.278, $p < .01$], *Initial Pause* [β = 0.203, $p < .05$], and *0.5 Second Pauses* [β = 0.208, $p = .06$]. The cross validation model was significant, $R^2$ = .204, *RMSE* = 0.187.

**Syntactic Complexity.** None of the keystroke indices were significantly or marginally significantly correlated with the selected measure of syntactic complexity.

**Semantic Diversity.** The analysis to predict the semantic diversity in essays yielded a significant model, $R^2$ = .375, *RMSE* = 0.123, $p < .001$. Five variables were significant predictors in this regression analysis: *1 Second Pauses* [β = -0.379, $p < .001$], *StdDev Events* (30s) [β = -0.361, $p < .01$], *Slope Degree* (30s) [β = 0.336, $p < .01$], *Median Latency* [β = -0.265, $p < .05$], and *Local Extremes* (60s) [β = 0.173, $p < .05$]. The cross-validation analysis yielded a significant model, $R^2$ = .255, *RMSE* = 0.133.

**Global Semantic Cohesion.** Analyses to predict global semantic cohesion based on keystroke data yielded a significant model, $R^2$ = .194, *RMSE* = 0.238, $p < .001$ with four significant predictors: *StdDev Events* (30s) [β = 0.477, $p < .01$], *3 Second Pauses* [β = 0.424, $p < .001$], *Verbosity* [β = 0.337, $p < .01$], and *Median Latency* [β = 0.307, $p < .05$]. The model produced by the cross-validation analysis was significant, $R^2$ = .160, *RMSE* = 0.244.

The results of the linguistic analyses indicate that the basic and time-sensitive keystroke indices were meaningfully related to the linguistic features of students' essays at multiple levels. Notably, however, the linguistic regression models were weaker than the essay score model, and the findings were less robust to the cross-validation procedure.

The model generated to predict semantic diversity was the strongest of the linguistic models. This analysis indicated that more semantically diverse essays were related to shorter pauses, with more variability at the 60-second time window (Local Extremes), but less variability at the 30-second time windows. The global semantic cohesion and word familiarity models were also significant with keystroke indices for both accounting for just under 20% of the variance in the linguistic properties. Finally, the syntactic complexity measure was not significantly related to any of the keystroke indices, indicating that perhaps behavioral patterns do not manifest in the different sentence structures produced by writers.

# 5. DISCUSSION

AWE systems provide an environment for students to receive writing instruction and engage in deliberate practice with summative and formative feedback [12]. Despite the general success of their scoring algorithms (e.g., [13-14; 27]), however, the pedagogical elements of these systems have much room for improvement. For instance, one major weakness of AWE systems is that they they typically only adapt to student users based on individual essay drafts. System developers tend to rely on NLP methods to examine the quality of students' written products; yet, information about their behavioral processes is largely ignored.

In the current study, we used system logs of keystrokes to develop online assessments of students' writing performance. The behavioral processes enacted by writers are important elements of writing skill [1; 22]; therefore, our aim was to determine whether we could assess and model the quality and linguistic properties of students' essays by calculating indices related to their typing behaviors. Basic and time-sensitive keystroke indices were calculated to analyze the behavioral patterns enacted by student writers. These indices provided information about writing processes at both the aggregate level (e.g., total number of pauses and backspaces) as well as information about how these behaviors unfolded over time. The results revealed that keystroke indices were able to model over three-quarters of the variance in students' essay scores. Additionally, these indices were able to model the linguistic properties of the essays at multiple levels.

The essay score analyses revealed that 10 keystroke indices were significantly correlated with students' holistic essay scores. This is important because it indicates that information about the quality of students' essays can be detected by analyzing their behavioral processes. Further, the two regression analyses revealed that the total number of keystrokes pressed by writers provided the most predictive power in the model, but that without this measure of Verbosity, the remaining indices were still about to account for 61% of the variance in essay scores.

These initial analyses of essay score indicate that *fluency* may be an important skill that is captured by the keystroke indices. In our study, the students who produced higher-quality essays were also more consistent in their typing (i.e., whether they typed or not) across the 30 second time windows, yet they had higher variability in the *number* of keystroke events they produced in these time windows. This finding suggests that these students' writing sessions may have been characterized by short (rather than long) patterns of writing and pausing. Some confirmation for this intuition is found in the the negative correlations between essay score and pause times (i.e., Median and Largest Latency). However, future research will need to examine these writing-pause patterns more closely. It may be the case, for instance, that short pauses are indicative of thoughtful writing, such as the search for appropriate words or phrases rather than "freewriting" behavior. Long pauses, on the other hand, may be indicative of mind wandering that warrants system intervention.

Follow-up linguistic analyses similarly revealed important information about the role of behavioral processes in writing. These analyses first indicated that the basic and time-sensitive keystroke indices were significantly related to the linguistic features of students' essays at the lexical, semantic and global cohesion levels, but not at the syntactic level. This indicates that keystroke indices may be picking up on specific meaning-making processes, rather than differences in cognitive factors, such as working memory capacity. For instance, semantic diversity represents the number of semantically related concepts that appear in students' essays, which may map onto the differences in the content that students chose to include in their essays. Syntactic complexity, on the other hand, is much more weakly related to the

*meaning* of a particular text and, instead, may be indicative of individual differences in specific cognitive skills (e.g., [31; 34]).

It is important to note that the keystroke indices accounted for a smaller amount of the variance in linguistic properties than in the overall essay scores. This suggests that variations in students' behavioral patterns may manifest in the properties of students' essays in different ways depending on the specific context. For instance, long pauses may be more indicative of cohesion if students are writing about an unfamiliar topic that requires more deliberate planning. On the other hand, if students are writing in response to a familiar or emotionally charged topic, it may be the case that essay cohesion will be associated with rapid typing with minimal pauses. The results of these follow-up analyses suggest that future analyses may need to use content-based information to make predictions about the relevance and interpretation of particular keystroke indices. Analytic techniques that allow the system to take past behavior and prompt content into consideration, for instance, could go a long way in improving the interpretability of these patterns.

These results are promising and suggest that keystroke indices can be utilized to uncover important information about the behavior and performance of student writers. Here, we analyzed the keystrokes produced for a short, prompt-based essay task. In the future, additional studies will be conducted to specifically examine how these keystroke patterns map onto writing across different genres, contexts, and difficulty levels. For example, multiple writing sessions could be collected for each participant, with prompt difficulty, genre, or audience varying across these sessions. This research design would help to disentangle signals that vary across multiple factors, such as boredom and difficulty.

Another area for future research lies in the calculation of more sophisticated keystroke indices, as well as the integration of keystroke indices with other system information. We used only keystroke indices as our predictors because we were interested in the degree to which simple behavioral measures alone could predict information about students' essays. In future studies, it will be important to consider additional indices that may be related to the context of these writing behaviors. For instance, if we aim to model students' engagement during writing, it will be important to collect additional information from our systems, such as their prior writing behaviors (e.g., on previous essays, or from original to revised drafts), as well as the linguistic content of the essays.

The overarching goal of this research is to enhance AWE systems such that they provide feedback and instruction that is more attuned to writers' processes. Eventually, we aim to be able to identify specific behavioral patterns associated with different writing processes, which will allow us to provide students with more pointed, online feedback and instruction. For example, through the combination of multiple keystroke indices, systems may be able to distinguish when students are experiencing writer's block as opposed to when they are engaged in the task, but have paused to think. If writer's block were detected, W-Pal could then ask students if they need help or offer specific strategies and practice opportunities for idea generation.

Overall, our results suggest that time-sensitive behavioral data can (and, in our opinion, should!) be used to help drive more personalized feedback and instruction in computer-based learning environments. Although a number of future studies are needed to investigate how this keystroke information can be used most effectively, the current study takes a strong first step in revealing the power of these indices.

## 6. ACKNOWLEDGEMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A120707 to Arizona State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

## 7. REFERENCES

[1] Graham, S. 2006. Writing. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (pp. 457-477). Mahwah, NJ: Erlbaum.

[2] National Commission on Writing. 2003. *The Neglected "R."* College Entrance Examination Board, New York.

[3] Allen, L. K., Snow, E. L., Crossley, S. A., Jackson, G. T., and McNamara, D. S. 2014. Reading comprehension components and their relation to the writing process. *L'année psychologique/Topics in Cognitive Psychology, 114*, 663-691.

[4] Allen, L. K., Snow, E. L., and McNamara, D. S. 2016. The narrative waltz: The role of flexibility on writing performance. *Journal of Educational Psychology*. doi: 10.1037/edu0000109

[5] Donovan, C. A., and Smolkin, L. B. 2006. Children's understanding of genre and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.) *Handbook of writing research* (pp. 131-143). New York: Guilford.

[6] Baer, J. D., and McGrath, D. 2007. *The reading literacy of U.S. fourth-grade students in an international context: Results from the 2001 and 2006 Progress in International Literacy Study (PIRLS)*. National Center for Educational Statistics, Institute of Education Sciences, U.S. Department of Education.

[7] National Assessment of Educational Progress. 2007. *The nation's report card: Writing 2007*. Retrieved Nov. 20, 2010, nces.ed.gov/nationsreportcard/writing/

[8] National Assessment of Educational Progress. 2011. *The nation's report card: Writing 2011*. Retrieved Nov. 5, 2012, nces.ed.gov/nationsreportcard/writing.

[9] Graham, S. and Perin, D. 2007. A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445-476.

[10] Kellogg, R., and Raulerson, B. 2007. Improving the writing skills of college students. *Psychonomic Bulletin and Review, 14*, 237-242.

[11] Johnstone, K. M., Ashbaugh, H., and Warfield, T. D. 2002. Effects of repeated practice and contextual-writing experiences on college students' writing skills. *Journal of Educational Psychology, 94(2)*, 305-315.

[12] Allen, L. K., Jacovina, M. E., and McNamara, D. S. 2016. Computer-based writing instruction. In C. A. MacArthur, S.

Graham, & J. Fitzgerald (Eds.), *Handbook of writing research (2nd ed.)* (pp. 316-329). New York, NY: The Guilford Press.

[13] Dikli, S. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*.

[14] Shermis, M., and Burstein, J. (Eds.). 2003. *Automated essay scoring: A cross-disciplinary perspective*. Erlbaum, Mahwah, NJ.

[15] Allen, L. K., Crossley, S. A., Snow, E. L., and McNamara, D. S. 2014. Game-based writing strategy tutoring for second language learners: Game enjoyment as a key to engagement. *Language Learning and Technology, 18*, 124-150.

[16] Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., and McNamara, D. S. 2014. The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*, 39-59.

[17] Deane, P. 2013. On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*, 7–24.

[18] Perelman, L. 2012. Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121–131). Fort Collins, Colorado: WAC Clearinghouse/Anderson, SC: Parlor Press.

[19] Shute, V. J. 2011. Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

[20] Snow, E. L., Jacovina, M. E., Allen, L. K., Dai, J., and McNamara, D. S. 2014. Entropy: A stealth assessment of agency in learning environments. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining*, (pp. 241-244). London, UK.

[21] Bixler, R. and D'Mello, S. 2013. Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (pp. 225-234). New York, NY: ACM.

[22] Leijten, M., and Van Waes, L. 2013. Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*, 358-392.

[23] Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S. A., D'Mello, S. K., and McNamara, D. S. in press. Investigating boredom and engagement during writing using multiple sources of information: The essay, the writer, and keystrokes. In *Proceedings of the 6th International Learning Analytics and Knowledge (LAK) Conference.*

[24] Roscoe, R. D., and McNamara, D. S. 2013. Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105,* 1010-1025.

[25] Allen, L. K., Crossley, S. A., Snow, E. L., Jacovina, M. E., Perret, C. A., and McNamara, D. S. 2015. Am I wrong or am I right? Gains in monitoring accuracy in an intelligent tutoring system for writing. In A. Mitrovic, F. Verdejo, C. Conati, & N. Heffernan (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education*. Madrid, Spain.

[26] Roscoe, R. D., Snow, E. L., Allen, L. K., and McNamara, D. S. 2015. Automated detection of essay revising patterns: application for intelligent feedback in a writing tutor. *Technology, Instruction, Cognition, and Learning, 10,* 59-79.

[27] McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., and Dai, J. 2015. Hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35-59.

[28] McNamara, D. S., Graesser, A. C., McCarthy, P., and Cai, Z. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University.

[29] McNamara, D. S., Crossley, S. A., and Roscoe, R. D. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45*, 499-515.

[30] Haberlandt, K., and Graesser, A. C. 1985. Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology, 114,* 357-374.

[31] Allen, L. K., Perret, C., and McNamara, D. S. in press. Linguistic signatures of cognitive processes during writing. Manuscript submitted to the *Annual Cognitive Science (Cog Sci) Society conference*.

[32] Landauer, T., McNamara, D. S., Dennis, S., and Kintsch, W. (Eds.). 2007. *Handbook of Latent Semantic Analysis.* Mahwah, NJ: Erlbaum.

[33] Chenoweth, N. A., and Hayes, J. R. 2001. Fluency in writing generating text in L1 and L2. *Written communication, 18*(1), 80-98.

[34] Kemper, S., Rash, S., Kynette, D., and Norman, S. 1990. Telling stories: The structure of adults' narratives. *European Journal of Cognitive Psychology*, *2*, 205-228.