Investigating Boredom and Engagement during Writing Using Multiple Sources of Information: The Essay, The Writer, and Keystrokes

Laura K. Allen
Arizona State University
PO Box 872111
Tempe, AZ, 85287
LauraKAllen@asu.edu

Caitlin Mills
University of Notre Dame
118 Haggar Hall
Notre Dame, IN, 46556
cmills4@nd.edu

Matthew E. Jacovina
Arizona State University
PO Box 872111
Tempe, AZ, 85287
Matthew.Jacovina@asu.edu

Scott Crossley
Georgia State University
34 Peachtree Ave, St 1200
Atlanta, GA 30303
scrossley@gsu.edu

Sidney D'Mello
University of Notre Dame
118 Haggar Hall
Notre Dame, IN, 46556
sdmello@nd.edu

Danielle S. McNamara
Arizona State University
PO Box 872111
Tempe, AZ, 85287
Danielle.McNamara@asu.edu

ABSTRACT

Writing training systems have been developed to provide students with instruction and deliberate practice on their writing. Although generally successful in providing accurate scores, a common criticism of these systems is their lack of personalization and adaptive instruction. In particular, these systems tend to place the strongest emphasis on delivering accurate scores, and therefore, tend to overlook additional indices that may contribute to students' success, such as their affective states during writing practice. This study takes an initial step toward addressing this gap by building a predictive model of students' affect using information that can potentially be collected by computer systems. We used individual difference measures, text indices, and keystroke analyses to predict engagement and boredom in 132 writing sessions. The results suggest that these three categories of indices were successful in modeling students' affective states during writing. Taken together, indices related to students' academic abilities, text properties, and keystroke logs were able classify high and low engagement and boredom in writing sessions with accuracies between 76.5% and 77.3%. These results suggest that information readily available in writing training systems can inform affect detectors and ultimately improve student models within intelligent tutoring systems.

Categories and Subject Descriptors

K.3.1 **[Computer Uses in Education]** *Computer-assisted instruction (CAI)*; I.2.7 **[Natural Language Processing]** *Text analysis, discourse*; J.5 [**Computer Applications: Arts and Humanities]**: Linguistics

General Terms

Algorithms, Measurement, Performance, Languages, Theory

Keywords

Intelligent Tutoring Systems, Natural Language Processing, stealth assessment, corpus linguistics, writing

1 INTRODUCTION

An individual's ability to effectively communicate ideas through text is an increasingly important skill in today's society. Indeed, in both educational and professional contexts, writing skills have become necessary for success [1-2]. Unfortunately, strong writing skills can be extremely challenging for students to develop and refine. This is largely due to the complex host of skills required to produce high-quality texts, such as strategically managing memory, developing strong vocabulary knowledge, setting goals, and producing coherent arguments [3-4]. Given the difficulty of developing these skills, it is not surprising that students consistently underachieve on tests of writing proficiency [e.g., 5-6].

In order for students to successfully develop the skills needed to produce high-quality texts, they need to be provided with explicit instruction and feedback. Specifically, research on writing instruction suggests that students benefit most from a combination of strategy instruction [7] and extended practice with individualized feedback [8]. One significant problem with these recommendations, however, relates to the difficulty of implementing them within typical classrooms. The time needed to prepare classroom materials, teach courses, and read, edit, and provide personalized feedback on students' essays can be overwhelming for teachers. This is particularly true today, as reports indicate that teachers are now faced with increasingly large

LAK '16, April 25-29, 2016, Edinburgh, United Kingdom
© 2016 ACM. ISBN 978-1-4503-4190-5/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2883851.2883939

class sizes and, as a result, have less time to devote to instruction, planning, and grading [9].

To help alleviate some of the difficulties facing writing instructors, researchers and technology developers have placed an increased focus on designing computer-based systems that can provide students with automated writing instruction and practice [10-11]. These writing training systems have been developed with a number of different goals in mind, ranging from the automatic scoring of essays to the delivery of personalized feedback and the instruction of writing strategies [12-15]. Automated essay scoring (AES) systems, for instance, focus on the *assessment* of the structure, content, and quality of student essays [11; 16]. These systems largely rely on natural language processing (NLP) and machine learning techniques to accurately model the scores assigned by expert raters [17].

These standalone AES systems have more recently been incorporated into educational learning environments, such as automated writing evaluation (AWE) systems [18-21] and intelligent tutoring systems (ITSs) [13]. The goal of these systems extends beyond the assessment of essay quality – rather, they focus on providing students with personalized feedback, as well as (in some cases) explicit instruction.

Despite their general success [e.g., 11-14], however, these systems have not gone without criticism [e.g., 16; 22-24]. In particular, critics have noted that AES assessments often miss out on components of rhetorical effectiveness and argumentation, and the feedback can often be impersonal and lacking in human sensitivity. The concerns noted about these systems are valid, and pose new challenges for developers of writing training systems. In particular, researchers have begun to shift their focus from simply providing *accurate* essay scores to providing feedback and adaptive instruction that is more nuanced and focuses on specific characteristics of individual students.

To illustrate the importance of this goal, consider two students, Kevin and Cecile, who write and submit essays to a particular writing system. While Kevin may be deeply engaged and interested in the topic of the essay prompt, his focus on the content might cause him to lose sight of some of the details that could improve his essay score, such as choosing more appropriate words and correcting spelling errors. Cecile, on the other hand, may produce an essay that is generally lacking in basic errors; however, her disinterest and boredom in the assignment may be apparent in her lack of compelling arguments and "attention grabbing" techniques. In this example, both students receive the same score from the system; however, their different affective states while writing the essay suggests that they may benefit from different feedback and adaptive instruction. Kevin may benefit from targeted feedback that acknowledges his effort and investment in the task, but that encourages him to take an additional look at the essay to improve grammar errors and word choices. Cecile, on the other hand, may benefit from feedback that reminds her of the importance of the topic, or suggests she engage in a game-based practice activity to increase her motivation.

One way to adjust to these differences among students and learning sessions is to embed assessments that are based on more than their essay scores. These measures can be hidden from users (i.e., "stealth assessments" [25-26]) and can inform more specific instruction and feedback that is tailored to students' strengths and weaknesses, as well as their potential affective states and learning preferences. Recent research suggests that affect is present through the writing process, and different affective states can predict the quality of students' writing outcomes [27-28]. There has been some success with respect to detecting affective states in computer-based learning environments (e.g., as reviewed in [29-30]); however, limited attention has been paid to developing affect detectors for systems that develop writing proficiency.

In the current paper, we address this gap in the literature by examining the efficacy of indices commonly collected in writing training systems to detect students' affective states during individual writing sessions. In particular, we examine whether individual difference measures, linguistic and semantic properties of the generated text, and keystroke measures can be used to model affective states. Second, we aim to determine whether each of these index types (i.e., individual differences, text properties, and keystroke measures) contribute unique predictive power in modeling affect during writing. Our ultimate goal is to use these models to provide more individualized tutoring and feedback to students.

1.1 Adaptive Feedback and Instruction

In an effort to provide more adaptive instruction and feedback to students, computer-based learning environments often rely on measures of performance (and other relevant indices) that can be collected without disrupting the learning task [25-26]. These "stealth assessments" can take many forms, from the trajectories of a user's mouse movements to the linguistic structure of their text responses. Most importantly, once these assessments have been developed, they can be used to improve student models, which can inform feedback delivery and instructional recommendations [31].

1.1.1 Individual Difference Measures

One potential method for increasing the validity and personalization of AWE systems (and other computer-based learning environments) is to first take into consideration any information that is already known about the student users. Given that one of the strongest criticisms facing developers of writing training systems is that the systems are impersonal, the inclusion of student-level information, such as literacy skills and cognitive abilities, may increase the sensitivity of algorithms to model student users. Indeed, relevant to this study, individual differences have been shown to be important predictors of affective states during writing [27]. This process of contextualizing the writing assessment based on individual differences is an important step, given that a primary goal of systems is to provide more adaptive and personalized feedback to students.

Recently, researchers have begun to consider the inclusion of individual differences in algorithms for predicting essay *score* [32]. In particular, Crossley and colleagues investigated the efficacy of improving traditional AES methods (i.e., statistical modeling human scores based on linguistic essay indices) by incorporating student-level indices into the model. The results of their study indicated that the combination of text and student indices led to scoring accuracies that were comparable to the industry-standard AES systems. Given the results of this study, it may be reasonable to assume that individual differences among students may similarly contribute to the accuracy of algorithms designed to measure student *affect*, rather than their essay scores.

1.1.2 Natural Language Processing and Writing

Indices related to text indices from the essays provide additional sources of information. Natural Language Processing (NLP) techniques are commonly employed in AES systems in order to extract various linguistic and semantic indices of students' essays. These indices have been used extensively in prior research on writing, particularly with the aim of improving models for predicting expert ratings of essay quality [11; 16; 33-35].

More recently, researchers have begun to investigate whether these NLP techniques can similarly be used to model individual differences among students. Allen and McNamara (2015) [36], for example, used indices related to the lexical properties of students' essays to successfully model their scores on an unrelated vocabulary knowledge assessment. Overall, these (and other) previous studies suggest that NLP techniques are an extremely powerful source of student data and can be used to inform stealth assessments to improve student models. Despite the wealth of previous research in this area, however, there is, to our knowledge, no current research testing the efficacy of these text indices to predict students' affective states during writing.

1.1.3 Keystroke Analyses for Writing

A final source of system data that may be useful for modeling students' affective states is the keystroke data related to the physical process of writing. Although researchers have made a significant effort to leverage the *indices* of texts to better understand writing quality and individual differences (as reviewed above), there has been significantly less research on students' "online" writing processes. Specifically, most of the previous research on writing has focused on students' finished writing products and not the moment-by-moment writing process. Studying the writing process can help to reveal qualities of a writer and written texts that are more difficult to measure with product measures alone. Keystroke logging tools have been developed to record the keys that writers press while typing. These tools provide a unique means to study the processes associated with writing [37-38], including investigations of struggling and expert writers [39] and a preliminary study on the detection on affective states during writing [37]. These tools are consistently improving; for example, InputLog, a prominent logging tool, can interface with NLP tools, affording analyses that include both keystroke information and linguistic information, such as parts of speech [40].

To illustrate the potentially important value of these keystroke analyses, consider the process of entering a state of "flow" during writing [41]. How might your patterns in keystroke timing vary when you enter into this flow state, as compared to when you are struggling to generate ideas or are bored? These differences may play a key role in the modeling students' affective states beyond the written text itself. Additionally, these indices may be able to help researchers identify and better understand the various states of productivity during writing, which can ultimately inform personalized feedback and instructional adaptations.

In a recent study, Bixler and D'Mello (2013) [37] conducted an initial investigation of these questions. In particular, they collected individual difference measures and keystroke data from student writers to detect on-line affective states during writing (i.e., self-reported affective states in 15-second intervals). Results of their analyses indicated that the combination of these behavioral measures and student-level indices was able to detect boredom, engagement, and neutral states between 11% and 38% above baseline. Additionally, their results were able to generalize to new individuals.

1.2 Writing Pal

One aim of the current research is to improve the adaptability of the Writing Pal (W-Pal) system. W-Pal is an intelligent tutoring system (ITS) that was developed to deliver explicit writing strategy instruction and practice to high school and early college students [13]. In contrast to the majority of writing training systems (see [10] for a review), W-Pal strongly focuses on the teaching of strategies for high-quality writing, in addition to providing multiple forms of practice (i.e., strategy-specific practice and holistic essay writing practice).

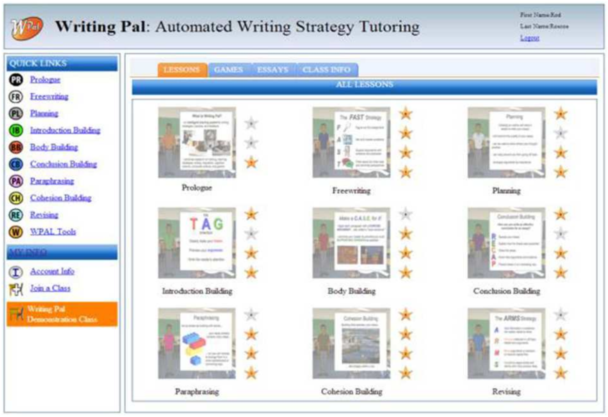
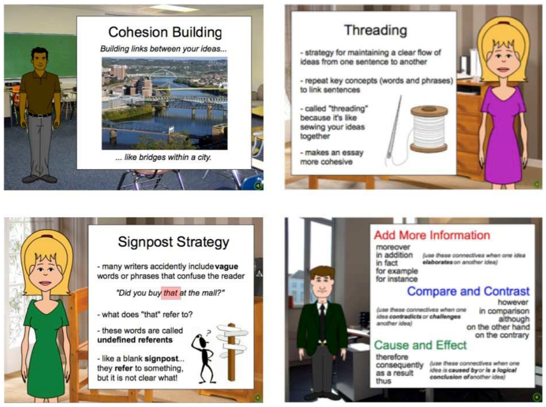


Figure 1: Main Interface of the W-Pal System

Strategy instruction in the W-Pal system covers the three primary phases of the writing process: prewriting, drafting, and revising. In the system, these strategies are taught in the context of individual instructional modules that include: Freewriting and Planning (prewriting); Introduction Building, Body Building, and Conclusion Building (drafting); and Paraphrasing, Cohesion Building, and Revising (revising; see Figure 1 for a screenshot of the main W-Pal interface). Each of these instructional modules contains multiple lesson videos, which are each narrated by an animated pedagogical agent (see Figure 2 for example screenshots of the videos) who describes and provides examples of specific strategies that are important for writing.


Cohesion Building

Building links between your ideas...

... like bridges within a city.

Threading

- strategy for maintaining a clear flow of ideas from one sentence to another
- repeat key concepts (words and phrases) to link sentences
- called "threading" because it's like sewing your ideas together
- makes an essay more cohesive

Signpost Strategy

- many writers accidently include **vague** words or phrases that confuse the reader

"Did you buy that at the mall?"

- what does "that" refer to?
- these words are called **undefined referents**
- like a blank **signpost**... they **refer** to something, but it is not clear what!

Add More Information

moreover
in addition
in fact
for example
for instance

*(use these connectives when one idea **elaborates** on another idea)*

Compare and Contrast

however
in comparison
although
on the other hand
on the contrary

*(use these connectives when one idea **contradicts** or **challenges** another idea)*

Cause and Effect

therefore
consequently
as a result
thus

*(use these connectives when one idea **is caused by** or **is a logical conclusion of** another idea)*

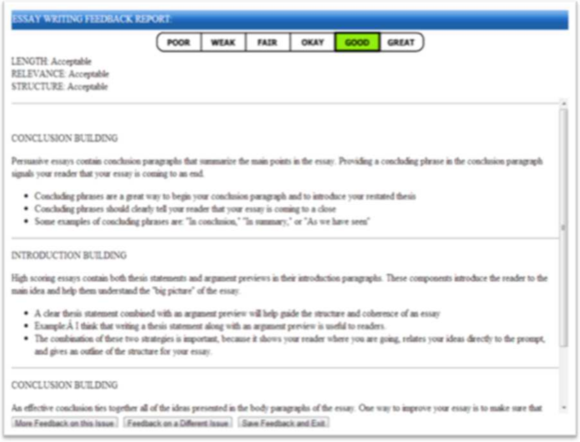
Figure 2: Screenshots of the W-Pal Lesson Videos

Once students have viewed the lesson videos, they can unlock mini-games that provide them with opportunities to practice the writing strategies in isolation before applying them in the context of a complete essay. In W-Pal, students can practice the strategies with identification mini-games, where they are asked to select the best answer to a particular question, or generative mini-games, where they produce natural language (typed) responses related to the strategies they are practicing.

1.2.1 W-Pal Essay Scoring and Feedback

An important component of the W-Pal system is the automated writing evaluation (AWE) component (i.e., the essay practice component). This aspect of W-Pal contains a word processor in which students can write essays in response to a set of SAT-style prompts. Additionally, teachers have the option of adding their own prompts to the system. Once a student has completed an essay, it is submitted to the W-Pal system for grading. The W-Pal algorithm [33] then calculates a variety of linguistic indices related to the submitted essay and provides both summative and formative feedback to the student (see Figure 3 for a screenshot of the feedback screen).

The summative feedback provided by W-Pal consists of a holistic essay score that ranges from 1 to 6 (described to students as "Poor" to "Great"). The formative feedback, on the other hand, provides information about the writing strategies that students can use to improve the quality of their essays. After they have read the feedback messages, students have the option to revise their essays based on the feedback that they received.



ESSAY WRITING FEEDBACK REPORT

LENGTH: Acceptable
RELEVANCE: Acceptable
STRUCTURE: Acceptable

CONCLUSION BUILDING

Persuasive essays contain conclusion paragraphs that summarize the main points in the essay. Providing a concluding phrase in the conclusion paragraph signals your reader that your essay is coming to an end.

- Concluding phrases are a great way to begin your conclusion paragraph and to introduce your restated thesis
- Concluding phrases should clearly tell your reader that your essay is coming to a close
- Some examples of concluding phrases are: "In conclusion," "In summary," or "As we have seen"

INTRODUCTION BUILDING

High scoring essays contain both thesis statements and argument previews in their introduction paragraphs. These components introduce the reader to the main idea and help them understand the "big picture" of the essay.

- A clear thesis statement combined with an argument preview will help guide the structure and coherence of an essay
- Example:Â I think that writing a thesis statement along with an argument preview is useful to readers.
- The combination of these two strategies is important, because it shows your reader where you are going, relates your ideas directly to the prompt, and gives an outline of the structure for your essay.

CONCLUSION BUILDING

An effective conclusion ties together all of the ideas presented in the body paragraphs of the essay. One way to improve your essay is to make sure that

Figure 3: Screenshot of the W-Pal Feedback

Formative feedback is an important component of writing development, as it provides important knowledge to writers about components of high-quality writing, as well as actionable recommendations for how to improve. Examples of these recommendations include: generating ideas and examples, maintaining cohesion through explicit text connections, and employing sophisticated words. The automated formative feedback in W-Pal was specifically developed with this in mind, and provides recommendations that relate to multiple writing strategies.

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 [42-43].

1.3 Current Study

The purpose of the current study is to investigate the degree to which the affective states that students experience during writing sessions can be classified based on measures readily collected by the W-Pal system. In particular, our aim is to use individual difference measures, text indices (e.g., linguistic and semantic text properties), and keystroke measures to classify whether a student experienced either high or low general levels of boredom and engagement over the course of an entire writing session. The overarching aim of this line of research is to develop stealth assessments of students' affective states during the writing process, which will ideally help to update student models in the W-Pal system. Increasing the sensitivity of W-Pal to students' affect is expected to improve its adaptability through the development of more nuanced and personalized feedback and recommendations.

To accomplish our initial goal, we collected essays from undergraduate students, along with a number of individual difference measures. Students provided retrospective judgments of their affective states during the writing process by viewing a video that displayed the student's face along with a screen-capture video of the computer interface used to write the essay. The linguistic and semantic properties of the essays were calculated using two NLP tools, Coh-Metrix [44], and SEANCE. Coh-Metrix calculates information related to linguistic indices of text, whereas SEANCE provides information related to semantic and affective information. Finally, we recorded the keystrokes logged during the writing process and calculated indices related to specific aspects of these keystrokes. We hypothesized that the individual differences, text properties, and keystroke indices would all provide unique predictive power in classifying students' writing sessions as high or low engagement and high or low boredom. Additionally, we predicted that engagement and boredom would be best classified by different combinations of indices. For instance, general engagement over the course of a writing assignment might be a more "fleeting" affective state, detectable by rapid bursts of activity, whereas boredom might manifest in more stable individual difference measures, such as a general aversion to writing.

2 METHODS

2.1 Participants

Participants were 44 undergraduate students from a university in the United States. Of these students, 68% were female, 45% were Caucasian, 52% were African American, and 3% reported "Other." The students reported a mean age of 19.9 years. All students participated in the study for course credit.

2.2 Individual Difference Measures

Participants were asked to self-report their ACT scores as a measure of their scholastic aptitude. Additionally, their apprehension towards writing was assessed with the Writing Apprehension test (WAT) [45]. The WAT is a 26-item self-report survey that prompts students to respond to multiple questions on a 5-point Likert scale related to their feelings toward the writing process. The WAT scores are negatively related to apprehension levels; thus, lower scores are indicative of more writing apprehension. Students' "exposure to print" was assessed using the Author Recognition Test [46]. Participants were shown a list of 42 popular authors, such as J.R.R. Tolkien or Dean Koontz) and were asked to check each author they recognized. Students' scores were simply the number of authors that were correctly recognized.

2.3 Data Collection Procedure

The participants were allotted 10 minutes to complete an essay on each of the three essay topics. For each, the students were first asked to select one of the subtopics described above. The participants typed their essays on a computer where each keystroke was logged, along with a timestamp, and the number of milliseconds that had passed since the last keystroke. Video of participants' faces and computer screens were also recorded. In all, participants completed three essays on the three topics (30 mins writing time total).

2.3.1 Retrospective Affect Judgment

The participants provided self-reports of their affective states immediately following the writing sessions. The judgments of the writing sessions began by playing a video of the participants' face along with their screen capture video on a computer monitor, similar to a cued-recall procedure [47-48]. The screen capture video included the writing prompt and dynamically presented the text as it had been written by the participants in order to provide them with the context of the writing session. Participants were instructed to make judgments on the affective states that were present at any moment during the writing session by manually pausing the videos. Additionally, they were instructed to make affect judgments at each 15-second interval – in these instances, the videos were automatically paused. Participants provided their judgments on a computer interface that allowed them to select one out of 15 affective states from an alphabetized drop down list. These states included: anger, anxious, boredom, confusion, contempt, curiosity, delight, disgust, fear, flow, frustration, happiness, neutral, sadness, and surprise. Altogether, these affective judgments were made based on the participants' facial expressions, contextual cues from the screen capture, the definitions of the affective states (presented on a piece of paper), and their memories of the writing session.

The affect judgment task yielded 5,551 affect judgments across the 44 participants. The fourteen affective states cumulatively accounted for 78.9% of the judgments, and neutral was reported for the remaining 21.1% of the judgments. Importantly, the most frequent affective state reported was *engagement (flow)* with an occurrence rate of 35.4%, followed by boredom at 26.4%. Together these two states accounted for over half of the affective observations. In the current study, we chose to focus on engagement and boredom because they comprised the majority of the observations, and the remaining affective states were either reported at very low frequencies or were inconsistently reported across participants. Boredom and engagement were also found to predict essay quality in previous research [28].

2.4 Corpus

The current corpus consisted of 132 essays, which were collected from a previous experiment that examined the role of affect during the writing process [27]. The experiment had a repeated measures design that prompted students to write essays on three different topics: *academic*, *socially charged*, and *personal/emotional experiences*. The order of the essay topics was counterbalanced across students using a 3 × 3 Latin Square design.

Participants were allowed to choose the "subtopic" of their essays from a list of options in order to maximize their engagement in the writing. The "academic" essay topics were adapted from the ACT test (standardized test in the U.S.) and the subtopics included: time spent in high school, the use of class discussions, and social skills that are taught in schools. The "socially charged" essay subtopics related to: abortion, gays in the military, and the death penalty. Finally, the subtopics for the "personal/emotional experience" essays included writing about an intense experience involving one of the six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise).

2.5 Essay Scoring

Two researchers scored the essays using a modified version of the SAT rubric [49]. The SAT is a standardized test commonly used for college admission in the United States. Essay quality was measured on a 6-point scale with a score of 1 indicating little to no mastery and several major flaws, 3 indicative a development of mastery, but with one or more major flaws, and a 6 indicating clear and consistent mastery with minor errors. The essays were randomly divided between the two raters who independently scored the entire corpus of essays. The resulting scores were then standardized within each rater in order to remove potential rater biases. Interrater reliability, computed on a random subset of essays scored by both raters, was $r = .91$.

2.6 Text Analyses

Linguistic and semantic indices of students' essays were obtained from component scores reported by the Coh-Metrix and SEANCE tools as discussed in greater detail below. In addition, the *total number of words* was computed for each essay, as this index is a strong predictor of essay quality [33].

2.6.1 Coh-Metrix

Coh-Metrix [44] is a computational text analysis tool that was developed, in part, to provide deeper measures of text difficulty. This tool analyzes texts at the word, sentence, and discourse levels; thus, it can potentially offer more information about the specific difficulties of a particular text. Previous work with Coh-Metrix suggests that multiple aspects of a text coordinate to affect subsequent comprehension. To account for these multiple textual aspects, Graesser and colleagues (2011) [50] developed the *Coh-Metrix Easability Components*. These components provide measures of the principal sources of text difficulty and are well aligned with an existing multilevel framework [51].

Narrativity. The narrativity of a text reflects the degree to which a story is being told, using characters, places, events, and other things familiar to readers. Highly narrative texts are typically easier to read.

Syntactic Simplicity. Syntactically simple texts contain shorter sentences and more familiar and simple syntax. These texts are typically easier to comprehend.

Word Concreteness. This component refers to texts that contain concrete and meaningful words that can easily evoke mental images. Increases in word concreteness correspond to easier and more understandable texts.

Referential Cohesion. Referential cohesion reflects the degree to which words and ideas overlap across a text. Texts that are high in referential cohesion represent explicit connections between ideas and are, consequently, easier to read.

Deep Cohesion. Deep cohesion refers to the presence of causal, intentional, and temporal connectives in a text. Texts with more deep cohesion afford readers to form strong representations of causal events and are typically easier to comprehend.

2.6.2 SEANCE

The SEntiment ANalysis and Cognition Engine (SEANCE) is a sentiment analysis tool that relies on a number of pre-existing sentiment, social positioning, and cognition dictionaries. Unlike other sentiment analysis tools commonly used in learning analytic studies (i.e., LIWC) [52], SEANCE is freely available and contains part of speech (POS) tags and valence indices. The tool is available at http://www.kristopherkyle.com/seance.html.

SEANCE indices are taken from available source databases such as SenticNet and EmoLex. For many of these dictionaries, SEANCE provides a negation feature (i.e., a contextual valence shifter) that ignores positive terms that are negated. The negation feature, which is based on Hutto and Gilbert [53], checks for negation words in the three words preceding a target word.

SEANCE also includes the Stanford part of speech (POS) tagger [41] included in Stanford CoreNLP. The POS tagger allows for POS tagged specific indices for nouns, verbs, and adjectives. POS tagging is an important component of sentiment analysis because unique aspects of sentiment may reside more strongly in adjectives or in verbs and adverbs.

The SEANCE tool can report on almost 3,000 indices, but because such a large number of indices can be unwieldy, SEANCE also reports on 20 components derived from the SEANCE indices: *negative adjectives, social order, action, positive adjectives, joy, affect for friends and family, fear and disgust, politeness, polarity nouns, polarity verbs, virtue adverbs, positive nouns, respect, trust verbs, failure, well being, economy, certainty, positive verbs,* and *objects*. We focus on the scores from these components.

2.7 Text Analyses

To assess whether students' *online* writing behaviors (i.e., their keystroke patterns) were related to their self-reported affective states, we calculated a number of keystroke indices. These indices are described in Table 1.

Table 1: Keystroke Indices

	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
Mean Latency	The mean of all the differences in time between keystrokes per essay (not including initial pause)
Median Latency	The median of all the differences in time between keystrokes per essay (not including initial pause)
Initial Pause	The length of the first pause of an essay writing session
0.5 Second Pauses	The number of pauses above .5 seconds and below 1 second
1 Second Pauses	The number of pauses above 1 second and below 1.5 seconds
1.5 Second Pauses	The number of pauses above 1.5 seconds and below 2 seconds
2 Second Pauses	The number of pauses above 2 seconds and below 3 seconds
3 Second Pauses	The number of pauses above 3 seconds

2.8 Statistical Analyses

Statistical analyses were conducted to investigate the ability of individual differences, text properties, and keystroke indices to classify students' affective states during writing. As mentioned in the "Retrospective Affect Judgment" section above, our analyses focused solely on classifying boredom and engagement. Thus, to determine whether an essay writing session was considered high or low in boredom and high or low in engagement, we first conducted a median split analysis on students' affect ratings. To do this, we calculated affect proportion scores per essay, such that the sum of all affect proportion scores per essay was one. We then classified essay writing sessions as high (mean = 0.306) or low (mean = 0.011) boredom, and high (mean= 0.623) or low (mean = 0.082) engagement based on a median split of the respective distributions. Note that we treated each individual essay writing session separately, rather than accounting for within-subject variability.

We next conducted a number of statistical analyses to determine which indices would best classify student affect. The indices were divided into student-level indices (i.e., ACT, WAT, and Author Recognition scores), essay indices (i.e., Coh-Metrix and SEANCE component scores; essay quality), and keystroke indices.

Visual inspections of the data were conducted to assess normal distributions. These inspections were followed by square root transformations to ensure that the data was normally distributed. Multicollinearity of the variables was assessed as pair-wise correlations $r > .90$. In the case that indices demonstrated multicollinearity, the index that correlated most strongly with the relevant affect proportion score was retained in the analysis.

We first conducted MANOVAs to identify which indices exhibited significant differences across the high and low boredom and engagement groups. The MANOVAs were followed by stepwise discriminant function analyses (DFAs). In the DFAs, we used only the indices that demonstrated significant differences between the high and low boredom and high and low engagement groups in the MANOVA. We first conducted the DFA analysis on the entire corpus, and then validated the model using leave-one-out-cross-validation (LOOCV). In LOOCV, one essay was removed from the corpus for each analysis and the remaining essays were used as the training set. We tested the accuracy of the DFA model by examining its ability to classify the omitted essay. The process was repeated until each essay was omitted once in the test set. This analysis therefore allowed us to test the model's classifications on an independent essay (i.e., data that is not in the training set). If results on training and testing on all essays (i.e., no separate test set) and the LOOCV set are similar, confidence in model stability is increased.

3 RESULTS

3.1 Boredom

A MANOVA was conducted comparing the differences in individual differences, text properties, and keystroke indices between essay writing sessions that were reported high and low in boredom. No two predictors correlated above $r = .90$; therefore, no indices were removed from the analysis.

The results of the MANOVA analysis indicated that 12 indices were significantly different across high and low boredom writing sessions (see Table 2 for descriptive statistics of the 12 indices).

The stepwise DFA retained six variables related to individual differences, essay properties, and keystroke indices: WAT scores, 3 Second Pauses, Narrativity Component Score, Polarity Noun Component, Number of Words, and Median Latency. The results revealed that the DFA using these six indices correctly allocated 102 of the 132 essays on the entire set, $\chi2$ (df=6, n=132)=57.27 p< .001, for an accuracy of 77.3% (the chance level for this analysis is 50%) For the LOOCV analysis, the DFA allocated 101 of the 132 essays for an accuracy of 76.5% (see the confusion matrix reported in Table 3 for results). It appears that students who reported more boredom during writing were also less likely to have apprehension towards writing. Additionally, the boredom ratings were related to a lower frequency of long pauses while writing, shorter pauses in general, and shorter essays that contained fewer narrative elements, but a higher number of nouns related to polarity.

Table 2: Descriptive statistics [Means and (SD)] for variables included in DFA

Variable	Low Boredom	High Boredom
Narrativity Component	85.80 (15.57)	78.68 (23.31)
WAT scores	68.98 (14.29)	58.38 (14.95)
Largest Latency	25995.92 (21834.62)	39233.18 (36521.24)
Median Latency	199.56 (59.85)	173.10 (39.20)
0.5 Second Pauses	114.91 (46.65)	87.94 (30.65)
1 Second Pauses	29.29 (10.97)	22.96 (9.17)
1.5 Second Pauses	12.69 (4.82)	10.49 (5.30)
3 Second Pauses	18.59 (6.42)	16.10 (5.33)
Number of Words	212.30 (85.87)	186.13 (62.39)
Action Component	65.20 (24.54)	56.30 (17.93)
Polarity Nouns	38.21 (33.89)	52.20 (41.05)
Trust Verbs	19.44 (13.89)	25.59 (13.94)

Table 3: Confusion matrix for DFA classifying low and high boredom

		Low Boredom	High Boredom
Whole Set	Low Boredom	**49**	15
	High Boredom	15	**53**
		Low Boredom	High Boredom
LOOCV	Low Boredom	**48**	15
	High Boredom	15	**53**

3.2 Engagement

Our second analysis examined the degree to which the indices could classify essay sessions as having high or low degrees of engagement/flow. A MANOVA was first conducted to determine which indices were significantly different across the high and low engagement essay sessions. This analysis yielded nine significant indices (see Table 4 for descriptive statistics of the 9 indices).

A stepwise DFA was calculated to investigate whether these nine indices accurately classified the writing sessions according to self-reported engagement. The resulting DFA model retained three variables: Author Recognition Test scores (+), Median Latency (-), and WAT scores (+). This model correctly allocated 101 of the 132 students in the total set, χ2 (df=3,n=132)=42.790 p< .001, for an accuracy of 76.5% (the chance level for this analysis is 50%). For the LOOCV analysis, the DFA allocated 96 of the 132 students for an accuracy of 72.7% (see the confusion matrix reported in Table 5). Thus, students who experienced a higher proportion of engagement during writing had a greater exposure to print, as well as less apprehension towards writing. These more engaged students had shorter pauses than writers who reported lower levels of engagement.

Table 4: Descriptive statistics [Means and (SD)] for variables included in DFA

Variable	Low Engagement	High Engagement
ACT Scores	20.12 (3.62)	22.25 (4.52)
WAT scores	58.46 (15.02)	68.43 (14.47)
Author Recognition Test scores	3.92 (1.76)	5.73 (2.52)
Overall Essay Z-Score	-.0.28 (0.85)	0.34 (1.00)
Verbosity	1306.62 (465.90)	1621.48 (636.94)
Backspaces	164.82 (104.07)	218.09 (153.72)
Mean Latency	440.27 (153.25)	358.04 (114.33)
Median Latency	205.84 (53.17)	166.62 (42.64)
Number of Words	179.32 (63.55)	217.73 (81.72)

Table 5: Confusion matrix for DFA classifying low and high engagement

		Low Engagement	High Engagement
Whole Set	Low Engagement	**54**	11
	High Engagement	20	**47**
		Low Boredom	High Boredom
LOOCV	Low Engagement	**51**	14
	High Engagement	22	**45**

4 DISCUSSION

Writing training systems have been developed to provide students with instruction and deliberate practice on their writing [10]. While generally successful in providing accurate summative feedback [11-12], a common criticism of these systems is their lack of personalization and adaptive instruction [22-24]. The objective of most writing training systems is to provide accurate scores that match an expert's ratings of the essay's quality. These systems tend to overlook additional variables that may ultimately contribute to students' success, such as their affective states during writing practice.

Teachers can observe and interpret students' affect before and after they compose a writing assignment. They can then use these judgements to guide what feedback they give, and how they convey that feedback. We believe that it is both possible and desirable for automated systems to do much the same. But accomplishing this goal will require a more comprehensive picture of the writer and the writing process. This study takes an initial step toward this goal by building a predictive model of students' affect using information that can potentially be easily collected by computer systems.

We used individual difference measures, text indices (calculated via NLP tools), and keystroke analyses to predict affect. The MANOVAs revealed that there were 12 indices that significantly differentiated between low and high boredom writing sessions, and 9 indices that significantly differentiated between the low and high engagement groups. This is an important finding because it indicates that students' affective ratings can be detected by analyzing information about the students, as well aspects of the final product (i.e., text indices) and the writing process (i.e., keystrokes). Further, the DFA analyses revealed that boredom ratings were predicted by all three categories of indices – namely, students who frequently reported feeling bored during the writing session reported higher levels of writing apprehension, wrote shorter and less narrative essays, and had a lower frequency of long pauses. Engagement, on the other hand, was largely characterized by student and process-level indices, such as lower writing apprehension and shorter pause lengths.

Importantly, these DFAs revealed both similarities and differences between the boredom and engagement ratings. First, both the high boredom and high engagement groups were classified by shorter pause lengths. This is an interesting finding and potentially suggests that high levels of boredom and engagement may have been co-present during specific writing sessions. In particular, nearly half (43%) of the writing sessions were categorized as having high boredom and high engagement (23%) or low boredom and low engagement (20%). Thus, there may have been specific students who experienced higher degrees of affect in general during their writing sessions, as opposed to reporting neutral affective states.

This "pause" finding points to multiple promising areas of future research. First, follow-up studies that examine individual differences may reveal student profiles associated with varying levels of engagement and boredom during their writing sessions. A second follow-up study relates to the detection of affect *during* the writing session. In the current study, we categorized whether an essay writing session had a high or low proportion of boredom and engagement ratings. However, future studies should focus on the development of affect detectors that can signal the system when it predicts a student is experiencing certain emotions that warrant feedback (e.g., boredom).

In addition to this similarity in pause times, the DFAs indicated some differences between the engagement and boredom ratings. While high levels of boredom were associated with lower apprehension and shorter, less narrative essays, high engagement was predicted by lower writing apprehension and author recognition scores. This suggests that students' feelings of boredom may be more strongly related to the text that they produce, as it is related to both text indices and keystroke indices; engagement, on the other hand, may be more strongly influenced by more stable traits of the students. This has important implications for future system adaptability. If this finding were to be replicated in follow-up studies, it suggests that boredom and engagement should potentially be addressed in different ways. Boredom, for instance, might require more "online" feedback, whereas low levels of engagement might be addressed through the assignment of more motivating prompts prior to the writing session.

As a final note, in the current study, we focused on students' individual writing sessions, and did not account for within-subject variability associated with students' multiple writing sessions.[1] This methodological choice was made because the majority of current AES systems focus on assessing writing quality at the individual session level and do not account for students' previous performance. Although it may be the case that students' prior performance and affective states can increase the strength of the feedback and instructional adaptation in these systems, this remains an empirical question. Future studies should be conducted to compare the effectiveness of AES systems that do and do not account for previous student performance in their models.

Overall, our results suggest that individual differences, text indices, and keystroke logs can be utilized to develop models of students' affective states during writing sessions. Taken together, indices related to students' academic abilities, text properties, and keystroke logs were able to reliably predict the general affective states that students experienced during writing. These results are important because they suggest that students' affect can manifest in the ways that they produce essays, both in the indices of the texts themselves, as well as in their typing patterns. In the current study, we focused solely on engagement and boredom. However, future studies will be conducted to examine additional affective states, as well as other individual differences that may help to improve the adaptability of writing training systems.

In conclusion, the current study utilized multiple components related to the writing process to investigate the efficacy of writing training systems to inform stealth assessments of students' affective states. Our eventual goal is to use these stealth assessments to enhance our student models in the W-Pal system, which will allow us to provide students with more personalized feedback and instruction. More broadly, the current study suggests that individual differences, text indices, and online writing measures (such as keystroke analyses) can be used as a step towards more adaptive educational technologies for writing. Although this is only a first step, and a number of studies remain to be conducted, this study provides a strong initial foundation because it demonstrates the feasibility of such measures for modeling affect.

5 ACKNOWLEDGMENTS

This research was supported in part by: NSF ITR 0325428, HCC 0834847, DRL 1235958, DRL 1417997, IIS 1523091, and IES R305A120707. Opinions, conclusions, or recommendations do not necessarily reflect the views of the IES or NSF. We also thank Rod Roscoe, Cecile Perret, and Jianmin Dai for their help with the data collection and analysis and developing the ideas found in this paper.

6 REFERENCES

[1] Geiser, S. and Studley, R. 2001. *UC and SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California*. Oakland, CA: University of California.

[1] To ensure that our models were not largely biased by within-subject variability, we conducted leave-one-subject-out cross-validation for the classification analyses of engagement and boredom. The most accurate classifications in each case came from a Bayes Net classifier with 70.1% accuracy for the Engagement model and a Logistic Regression with 70.4% for the Boredom model.

[2] Powell, P. 2009. Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication, 66,* 664-682.

[3] Flower, L. S. and Hayes, J. 1981. A cognitive process theory of writing. *College Composition and Communication, 32, 365-387.*

[4] Hayes, J. 1996. A new framework for understanding cognition and affect in writing. In C. M. Levy & L. S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications*. Erlbaum, Mahwah, NJ, 1-27.

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

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

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

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

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

[10] Allen, L. K., Jacovina, M. E., and McNamara, D. S. 2015. Computer-based writing instruction. In C. A. MacArthur, S. Graham, and J. Fitzgerald (Eds.), *Handbook of writing research (2nd ed.)* (pp. 316-329). New York: Guilford Press.

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

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

[13] 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.

[14] Weigle, S. C. 2013. English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 1,* 85–99.

[15] Xi, X. 2010. Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27,* 291–300.

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

[17] Warschauer, M., and Ware, P. 2006. Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*, 1–24.

[18] Attali, Y., and Burstein, J. 2006. Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4(3).* Retrieved from www.jtla.org

[19] Crossley, S. A., Varner, L. K., Roscoe, R. D., and McNamara, D. S. 2013. Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In: K. Yacef et al (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED)* (pp. 269-278). Springer, Heidelberg, Berlin, 269-278.

[20] Grimes, D., and Warschauer, M. 2010. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment, 8,* 4–43.

[21] Shermis, M. D., Burstein, J., Elliot, N., Miel, S., and Foltz, P. W. 2015. Automated writing evaluation: An expanding body of knowledge. In C. A. MacArthur, S. Graham, & J. Fitzgerald, *Handbook of writing research (2nd ed.).* New York: Guilford Press.

[22] Haswell, R. H. 2006. Automatons and automated scoring: Drudges, black boxes, and dei ex machina. In: P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 57–78). Logan, UT: Utah State University Press.

[23] Hearst, M. 2002. The debate on automated essay scoring. *Intelligent Systems and their Applications, IEEE*, *15*, 22-37.

[24] 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.

[25] 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.

[26] Shute, V. J., and Kim, Y. J. 2013. Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop (Eds.), *Handbook of research on educational communications and technology (4th ed.)* (pp. 311-323). Lawrence Erlbaum Associates, Taylor & Francis Group, New York, NY.

[27] D'Mello, S.K., and Mills, C. 2014. Emotions during emotional and non-emotional writing. *Motivation and Emotion, 38*, 140-156.

[28] Mills, C. and D'Mello, S. K. 2013. Emotions during writing about socially-charged issues: Effects of the (mis) alignment of personal positions with instructed positions. In: *Proceedings of 26th Florida Artificial Intelligence Research Society Conference* (pp. 509-514). Menlo Park, CA: AAAI Press.

[29] Baker, R., and Ocumpaugh, J. 2015. Interaction-based affect detection in educational software. In R. Calvo, S. D'Mello, J. Gratch & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 233-245). New York: Oxford University Press.

[30] D'Mello, S., and Graesser, A. 2015. Feeling, thinking, and computing with affect-aware learning technologies. In R. Calvo, S. D'Mello, J. Gratch & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 419-434). New York: Oxford University Press.

[31] Brusilovsky, P. 1994. The construction and application of student models in intelligent tutoring systems. *Journal of Computer and Systems Science International, 23,* 70-89.

[32] Crossley, S. A., Allen, L. K., Snow, E. L., and McNamara, D. S. 2015. Pssst...Textual features...There is more to

Automatic Essay Scoring than just you!. In J. Baron, G. Lynch, N. Maziarz, P. Blikstein, A. Merceron, & G. Siemens (Eds.), *Proceedings of the 5th International Learning Analytics & Knowledge Conference (LAK'15)* (pp. 203-207). Poughkeepsie, NY.

[33] 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.

[34] Varner, L. K., Roscoe, R. D., and McNamara, D. S. 2013. Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research, 5*, 35-59.

[35] Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., and McNamara, D. S. 2014. Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment*.

[36] Allen, L. K., and McNamara, D. S. 2015. You are your words: Modeling students' vocabulary knowledge with natural language processing. In: O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015).* Madrid, Spain.

[37] 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.

[38] 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.

[39] Van Waes, L., Leijten, M., Lindgren, E., and Wengelin, A. 2015. Keystroke logging in writing research: Analyzing online writing processes. In C. A. MacArthur, S. Graham, & J. Fitzgerald, *Handbook of writing research (2nd ed.).* New York: Guilford Press.

[40] Leijten, M., van Horenbeeck, E., and Van Waes, L. 2015. Analyzing writing process data: A linguistic perspective. In G. Cislaru (Ed.), *Writing(s) at the crossroads: The process/product interface.* Philadelphia: John Benjamins Publishing Company.

[41] Kellogg, R. T. 2006. Professional writing expertise. In K. A. Ericsson, N. Charness, P. J. Feltovich, and R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 389-402). Cambridge University Press.

[42] 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.

[43] 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.

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

[45] Daly, J., and Miller, M. 1975. The empirical development of an instrument to measure writing apprehension. *Research in the Teaching of English 9(3),* 242-249.

[46] Cunningham, A. E., and Stanovich, K. E. 1997. Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33(6),* 934-945.

[47] D'Mello, S. and Graesser, A. C. 2011. The half-life of cognitive-affective states during complex learning. *Cognition and Emotion, 25, 1299-1308.*

[48] Rosenberg, E. L. and Ekman, P. 1994. Coherence between expressive and experiential system of emotions. *Cognition and Emotion,* 8, 201-229.

[49] McNamara, D. S., Crossley, S. A., and McCarthy, P. M. 2010. Linguistic indices of writing quality. *Written Communication, 27,* 57-86.

[50] Graesser, A.C., McNamara, D.S., and Kulikowich, J.M. 2011. Coh Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223-234.

[51] Graesser, A.C. and McNamara, D.S. 2011. Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 2,* 371-398.

[52] Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [computer software]. Austin, TX.

[53] Hutto C, Gilbert E. 2014. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: *International AAAI conference on weblogs and social media,* (pp. 216–225).