You are Your Words: Modeling Students' Vocabulary Knowledge with Natural Language Processing Tools

Laura K. Allen
Tempe, AZ, USA
Arizona State University
LauraKAllen@asu.edu

Danielle S. McNamara
Tempe, AZ, USA
Arizona State University
Danielle.McNamara@asu.edu

ABSTRACT
The current study investigates the degree to which the lexical properties of students' essays can inform stealth assessments of their vocabulary knowledge. In particular, we used indices calculated with the natural language processing tool, TAALES, to predict students' performance on a measure of vocabulary knowledge. To this end, two corpora were collected which contained essays from early college and high school students, respectively. The lexical properties of these essays were then calculated using TAALES. The results of this study indicated that two of the linguistic indices were able to account for 44% of the variance in the college students' vocabulary knowledge scores. Additionally, the significant indices from this first corpus analysis were able to account for a significant portion of the variance in the high school students' vocabulary scores. Overall, these results suggest that natural language processing techniques can inform stealth assessments and help to improve student models within computer-based learning environments.

Keywords
Intelligent Tutoring Systems, writing, Natural Language Processing, feedback

1. INTRODUCTION
Writing is a complex cognitive and social process that is important for both academic and professional success [1]. As contemporary societies grow increasingly reliant on text sources to communicate ideas (e.g., emails, text messages, online reports, blogs), the importance of developing proficiency in this area is more important than ever. Unfortunately, acquiring writing skills is no simple task – as evidenced by the many students who underachieve each year on national and international assessments of writing proficiency [1, 2, 3, 4]. Indeed, this text production process is complex and relies on the development of both lower and higher-level knowledge and skills, ranging from a strong knowledge of vocabulary to the strategies necessary for tying their ideas together [5, 6, 7].

To develop the skills that are required to produce high-quality texts, students need to be provided with comprehensive instruction that targets their individual strengths and weaknesses. In particular, this instruction should explicitly describe and demonstrate the skills and strategies that will be necessary during each of the phases of the writing process. Additionally, it should offer students opportunities to receive summative and formative feedback on their work, while engaging in deliberate practice. This form of *deliberate* practice is an important factor in students' development of strong writing skills [8, 9], because it can promote self-regulation of the planning, generation, and reviewing processes [9]. Unfortunately, however, deliberate practice inherently relies on individualized writing feedback. This is often difficult for teachers to provide, as they are faced with large class sizes and do not have the time to provide thorough comments on every essay that a student writes.

As a result of these classroom needs, researchers have developed computer-based writing systems that can provide students with feedback on their writing [10]. These systems have been used for both classroom assignments and high-stakes writing assessments to ease the burden of individualized essay scoring [11]. Specifically, *automated essay scoring* (AES) systems evaluate the linguistic properties of students' essays to assign them holistic scores [12, 13]. These systems use a multitude of natural language processing (NLP) and machine learning methodologies to provide these essay scores, and previous research suggests that they are often comparable to human raters [11, 13, 14, 15].

To provide students with greater context for the scores on their essays, AES systems are commonly incorporated into educational learning environments, such as *automated writing evaluation* (AWE) systems [16, 17] and *intelligent tutoring systems* (ITSs) [18]. These systems not only provide students with summative feedback on their essays (i.e., holistic scores), they also provide formative feedback and writing instruction. In order to be successful, these systems must contain algorithms that can provide individualized feedback that is relevant to students' individual skills.

Importantly, these computer-based writing environments rely on linguistic features to assess the *quality* of the individual essays submitted to the systems. Although the scores are generally valid and reliable, the systems rarely consider student-level information (e.g., their knowledge, skills, or affect) when providing feedback based on these scores. This can pose critical problems when developing adaptive components for the systems. As an example, consider two students, Mary and John, who both write essays that receive holistic scores of "3" from an AWE system. While Mary is able to clearly argue her point in the thesis and topic sentences,

her essay is weakened by simplistic language and sentence constructions. John, on the other hand, employs sophisticated vocabulary and eloquent sentences throughout his essay; however, he does a poor job of explaining his position on the argument. In this example, both students received the same score from the system; however, their essays were affected by different student-level strengths and weaknesses. Mary may have suffered from lower vocabulary knowledge and general language skills, whereas John may not have developed adequate planning and organization strategies.

One way to accommodate these individual differences is to develop user models based on students' characteristics, beyond simply their scores on essays. These models can provide more specific instruction and feedback that are tailored to students' strengths and weaknesses. One individual difference that may be particularly important to consider in these student models is *vocabulary knowledge.* Previous studies have shown that vocabulary knowledge plays a major role in the writing process, as it is strongly correlated with the scores assigned to students' essays [5, 19]. In the current paper, we examine the efficacy of NLP techniques to inform stealth assessments of this knowledge. In particular, we examine whether the lexical properties of students' essays can accurately model their scores on a standardized measure of vocabulary knowledge. Ultimately, our aim is to use these measures to provide more individualized tutoring to student users.

1.1 Stealth Assessments

In order to provide a more personalized learning experience (e.g., individualized instruction and feedback), computer-based learning environments must rely on repeated assessments of performance as students interact with the system. These measures can provide important information about students' knowledge states and learning trajectories, which can help to increase the adaptivity of these systems. Despite the importance of these assessments, however, they are not particularly conducive to robust student learning. In particular, constantly exposing students to questionnaires and tests can disrupt their learning flow [20] and subsequently harm their performance on later tasks.

As a response to this assessment problem, researchers have placed an emphasis on the development of methods that can accumulate information about student users without persistently disrupting the learning task [20, 21]. In particular, researchers have proposed the development of *stealth assessments*. These assessments are intended to measure students' performance and knowledge without requiring any explicit testing. Typically, these stealth assessments are embedded within the learning task itself and, as a result, are not able to be detected by students [22].

Within the context of computer-based learning environments, these stealth assessments can be informed by a wealth of information that can be easily logged in the system. These data can range from the speed at which someone is typing to the trajectories of their mouse movements. Snow and colleagues (2014), for example, developed stealth assessments of agency within a reading comprehension tutoring system [23]. They found that students who exhibited more systematic patterns of behavior in the system produced higher quality self-explanations compared to students who were more disordered in their choice patterns. They stated that this measure of behavior patterns could serve as a stealth assessment of agency in adaptive learning environments. Overall, stealth assessments can serve as a viable solution to the assessment problem, as they can be informed by a wide variety of data types to model the characteristics of student users (e.g., their skills, attitudes, etc.) [23, 24].

Importantly, after they have been developed, these stealth assessments can be used to enhance student models. Models of students' performance and attitudes are typically embedded in ITSs as a means to provide more individualized instruction and feedback [25]. In these systems, student users are represented by continuously updating models that are representative of their own knowledge and performance in the system. Thus, once the system has the ability to reliably assess students' particular skill sets, it can adapt in precise ways that can enhance the overall efficacy of the instruction [26].

1.2 Natural Language Processing

Natural language processing (NLP) tools provide a means through which researchers can develop stealth assessments of student characteristics [24]. In addition, these tools can help researchers to investigate the relationships between individual differences and the learning process at a more fine-grained size. By calculating indices related to multiple levels of the text (e.g., lexical, syntactic, discourse), researchers can look beyond simple measures of holistic quality (i.e., essay scores) and begin to examine and model the components of the writing process more thoroughly [27]. These models of student performance can then allow researchers and educators to provide students with more effective instruction that specifically targets their individual needs.

Broadly, NLP involves the automated calculation of linguistic text features using a computer program (or programming language) [28]. Thus, the focus of NLP primarily rests on the use of computers to understand, process, and produce natural language text for the purpose of automating certain communicative acts (e.g., providing technical support) or for studying communicative processes (e.g., examining the linguistic properties of readable texts). This technique can serve as a powerful methodological approach for researchers who are interested in examining particular aspects of the writing process [27] or for many other domains in which students produce natural language.

Researchers have employed NLP techniques within a variety of domains and contexts for the purpose of developing a better understanding the learning process [7, 24, 29, 30, 31]. For example, Varner, Jackson and colleagues (2013) used NLP tools to calculate the extent to which students' self-explanations of complex science texts contained cohesive elements [31]. Results from this study indicated that better readers produced more cohesive self-explanations than less skilled readers, indicating that automated indices of cohesion could potentially serve as a proxy for the coherence of students' mental text representations. In another study, Graesser and colleagues (2011) developed multiple components of text readability using NLP tools [29]. These components related to different dimensions of text complexity, such as narrativity, concreteness, and referential cohesion. Through the use of NLP tools, these researchers were able to develop components that provide multidimensional information about texts and the specific properties that influence students' ability to comprehend these texts successfully.

1.2.1 NLP and Writing

With regards to the writing process, NLP can serve as a particularly beneficial tool, as it can provide explicit information about students' processes and performance on the learning task. Accordingly, these NLP techniques have been used in previous research on writing, primarily with the goal of modeling human ratings of text quality [14, 30, 32]. In one particular study, Crossley and McNamara (2011) examined the linguistic indices that were significantly related to quality ratings of timed, prompt-based essays. Results of this study revealed that higher quality essays contained more sophisticated language, greater lexical diversity, more complex sentence constructions, and less frequent words. In a similar analysis, Varner and colleagues (2013) investigated differences between the linguistic indices associated with teachers' ratings of essay quality and students' self-assessments of their own essays [30]. This analysis suggested that students were less systematic in their self-assessments than teachers, at least in relation to the linguistic characteristics of the essays. Additionally, students' ratings were related to different linguistic features than the essay ratings of their teachers.

Overall, the results of these (and many other) studies suggest that NLP can serve as a powerful resource with which researchers can model the writing process at a more fine-grained size. In particular, NLP tools can potentially help researchers to develop better models of the individual differences that are important to writing proficiency (e.g., vocabulary knowledge), as well as for any other domain in which students produce natural language.

1.3 The Writing Pal

The Writing Pal (W-Pal) is an intelligent tutoring system (ITS) that was designed to provide explicit writing strategy instruction and practice to high school and early college students [18, 33] Unlike typical AWE systems, W-Pal places a strong emphasis on the instruction of writing strategies, as well as multiple forms of practice (i.e., strategy-specific practice and holistic essay writing practice).

The strategy instruction in W-Pal covers all three phases of the writing process: prewriting, drafting, and revising. Within W-Pal, these strategies are taught in individual instructional modules, which 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. In these videos, the agent describes and provides examples of specific strategies that are important for writing.

After viewing these lesson videos, students unlock multiple mini-games, which allow them to practice the strategies in isolation before applying them to complete essays. Within the W-Pal system, students can engage 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 strategy they are practicing.

One of the key features of the W-Pal system is its AWE component (i.e., the essay practice component). This system contains a word processor where students can write essays in response to a number of SAT-style prompts (teachers also have the option of adding in their own prompts to assign to students). Once a student has completed an essay, it is submitted to the W-Pal system. The W-Pal algorithm [14] then calculates a number of linguistic features related to the essay and provides summative and formative feedback to the student (see Figure 2 for a screenshot of the W-Pal feedback screen). The summative feedback in W-Pal is a holistic essay score that ranges from 1 to 6. The formative feedback in W-Pal provides information about strategies that students can employ in order to improve their essays. Once they have read the feedback, students have the option to revise their essays based on the feedback that they were assigned.

Figure 1. Main Interface of the W-Pal System

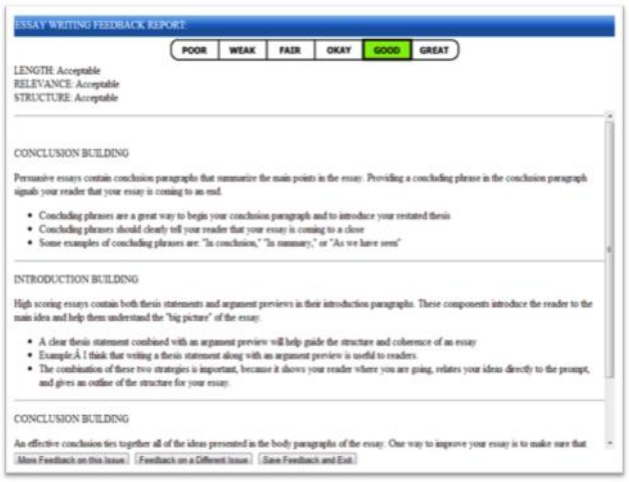

Figure 2. Example of W-Pal Feedback

2. CURRENT STUDY

The purpose of the current study is to investigate the degree to which the lexical properties of students' essays can inform stealth assessments of their vocabulary knowledge. Ideally, these assessments will serve to inform student models in the Writing Pal system and contribute to its adaptability in the form of more

sophisticated scoring algorithms, feedback, and adaptive instruction. To this end, two corpora were collected which contained essays from early college and high school students, respectively. The lexical properties of these essays were then calculated using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [34]. TAALES is an automated text analysis tool that provides linguistic indices related to the lexical sophistication of texts. We used this tool in the current study so that we could investigate the relationships between students' vocabulary knowledge and the lexical properties of the essays. We hypothesized that these lexical indices would be significantly related to vocabulary knowledge and that they would provide reliable measures of vocabulary knowledge across two distinct student populations.

2.1 Primary Corpus

The primary corpus for this study is comprised of 108 essays written by college students from a large university campus in Southwest United States. These students were, on average, 19.75 years of age (range: 18-37 years), with the majority of students reporting a grade level of college freshman or sophomores. Of the 108 students, 52.9% were male, 53.7% were Caucasian, 22.2% were Hispanic, 10.2% were Asian, 3.7% were African-American, and 9.3 % reported other ethnicities. All students wrote a timed (25-minute), prompt-based, persuasive essay that resembled what they would see on an SAT. Students were not allowed to proceed until the entire 25 minutes had elapsed. These essays contained an average of 410.44 words (*SD* = 152.50), ranging from a minimum of 84 words to a maximum of 984 words.

2.2 Vocabulary Knowledge Assessment

Students' vocabulary knowledge was assessed using the Gates-MacGinitie (4th ed.) reading comprehension test (form S) level 10/12 [35]. This assessment is a 10-minute task, which is comprised of 45 simple sentences that each contains an underlined vocabulary word. Students were asked to read each sentence and then select the most closely related word (from a list of five choices) to the underlined word within the sentence.

2.3 Text Analyses

To assess the lexical properties of students' essays, we utilized the Tool for the Automatic Analysis of Lexical Sophistication (TAALES). TAALES is an automated text analysis tool that computes 135 indices that correspond to five primary categories of lexical sophistication: *word frequency*, *range*, *n-gram frequencies*, *academic language*, and *psycholinguistic word information* [34]. These categories are discussed in greater detail below (see 34 for more thorough information).

Word frequency indices are indicative of lexical sophistication, because high frequency words are typically learned earlier in life, are processed more quickly, and are indicative of writing quality (i.e., with high frequency words indicating lower quality writing). There are two primary forms of frequency measures: frequency bands and frequency counts. Frequency bands measure the percentage of a text that occurs in particularly frequency bands (e.g., whether they are in the most frequent 1,000 words, 2,000 words in a frequency list, etc.). Frequency counts employ reference corpora and calculate the frequency of the words in a target text within the reference corpus.

Range indices are indicative of how widely used a particular word or family of words is. Thus, unlike frequency indices, range indices do not simply calculate a raw count of a word in a particular list or corpus. Rather, range indices measure the number of individual documents that contain that word in order to determine the extent that it is used broadly. Range has been used to successfully distinguish the frequent verbs produced by L2 speakers of English from the frequent verbs produced by native English speakers [36].

N-gram frequencies emphasize units of lexical items rather than single words. In particular, n-grams consist of combinations of *n* number of words (e.g., the bigram "years ago") that frequently occur together. Bigram lists have been shown to be predictive of a speaker or writer's native language, as well as the quality of a given text.

Academic language indices measure the degree to which a text contains words that are found infrequently in natural language corpora, but frequently in academic texts. A number of academic word lists have been calculated to measure the words that are commonly used in academic texts, such as textbooks and journal articles. Thus, these indices provide a measure of how academic a text is compared to more typical texts.

Psycholinguistic word indices provide information about the specific characteristics of the words used in texts. These properties have been shown to be related to lexical decision times, lexical proficiency, and writing quality. TAALES focuses on five particular properties of words: *concreteness* (i.e., perceptions of how abstract a word is), *familiarity* (i.e., judgments of how familiar words are to adults), *imageability* (i.e., judgments of how easy it is to imagine a word), *meaningfulness* (i.e., judgments of how related a word is to other words), and *age of acquisition* (i.e., judgments of the age at which a word is typically learned).

2.4 Statistical Analyses

Statistical analyses were conducted to investigate the role of lexical properties in assessing and modeling students' vocabulary knowledge scores. Pearson correlations were first calculated between students' scores on a vocabulary knowledge measure and the lexical properties of their essays (as assessed by TAALES). The indices that demonstrated a significant correlation with vocabulary knowledge scores ($p < .05$) were retained in the analysis. Multicollinearity of these variables was then assessed among the indices ($r > .90$). When two or more indices demonstrated multicollinearity, the index that correlated most strongly with vocabulary knowledge scores was retained in the analysis. All remaining indices were finally checked to ensure that they were normally distributed.

A stepwise regression analysis was conducted to assess which of the remaining lexical indices were most predictive of vocabulary knowledge. For this regression analysis, a training and test set approach was used (67% for the training set and 33% for the test set) in order to validate the analyses and ensure that the results could be generalized to a new data set. To additionally avoid overfitting the model, we chose a ratio of 15 essays to 1 predictor, which allowed 7 indices to be entered, given that there were 108 essays included in the analysis.

A final linear regression analysis was conducted to determine the extent to which these indices could model the vocabulary knowledge of students in a different population. In particular, we investigated whether the lexical sophistication indices that were retained in the previous regression model (i.e., the regression

model for the college students) accounted for a significant amount of the variance in a second set of students' (i.e., the high school students) vocabulary knowledge.

3. RESULTS

3.1 Vocabulary Knowledge Analysis for the Primary Corpus

Pearson correlations were calculated between the TAALES indices and students' Gates-MacGinitie vocabulary knowledge scores to examine the strength of the relationships among these variables. This correlation analysis revealed that there were 45 linguistic measures that demonstrated a significant relation with vocabulary knowledge scores and did not demonstrate multicollinearity with each other. To avoid overfitting the model, we only selected the 7 indices that were most strongly correlated with vocabulary knowledge. These 7 indices are listed in Table 1 (see Kyle & Crossley for explanations of each variable) [34].

A stepwise regression analysis was calculated with these 7 TAALES indices as the predictors of students' vocabulary knowledge scores for the students in the training set. This regression yielded a significant model, F (2, 76) = 29.296, p < .001, r = .660, R^2 = .435. Two variables were significant predictors in the regression analysis and combined to account for 44% of the variance in students' vocabulary knowledge scores: mean age of acquisition log score [β =.92, t(2, 76)=6.423, p < .001] and normed count for all academic word lists [β =-.36, t(2, 76)=-2.539, p = .013]. The regression model for the training set is presented in Table 2. The test set yielded r = .600, R^2 = .360, accounting for 36% of the variance in vocabulary knowledge scores.

Table 1. Correlations between Gates-MacGinitie vocabulary knowledge scores and TAALES linguistic scores

TAALES variable	*r*	*p*
Mean age of acquisition log score	.614	<.001
Mean range (number of documents that a word occurs in) log score	-.562	<.001
Spoken bigram proportion	-.511	<.001
Mean unigram concreteness score	-.492	<.001
Mean frequency score (bigrams)	-.488	<.001
Mean frequency log score	-.476	<.001
Normed count for all academic word lists	.402	<.001

Table 2. TAALES regression analysis predicting Gates-MacGinitie vocabulary knowledge scores

Entry	Variable added	R^2	Δ R^2
Entry 1	Mean age of acquisition log score	.387	.387
Entry 2	Normed count for all academic word lists	.435	.048

The results of this regression analysis indicate that the students with higher vocabulary scores produced essays that were more lexically sophisticated. The essays contained words that were acquired at a later age, such as the words *vociferous* or *ubiquitous*, which are predicted to be learned later than words such as *toy* and *animal*. The essays also contained a greater proportion of academic words that are frequently found in academic texts, such as *financier* or *contextualized*, rather than household words such as *bread* and *house*. Hence, better writers use words that are found in academic, written language, rather than more common, mundane language. Notably, these two indices, age of acquisition, and academic words, are likely to correlate with indices related to the frequency or familiarity of words in language. However, in this case, they more successfully captured students' vocabulary knowledge from their writing samples compared to simple frequency or familiarity indices.

3.2 Generalization to a New Data Set

Our second analysis specifically tested the ability of the linguistic indices to predict the Gates-MacGinitie vocabulary knowledge scores of students in a completely separate population. To address this question, we collected a test corpus of essays written by high school students and analyzed the lexical properties of these essays. Specifically, we calculated the *mean age of acquisition log score* and the *normed count for all academic word lists*, as these were the two indices retained in the previous regression model. These indices were then used as predictors in a regression model to predict students' vocabulary knowledge.

3.3 Test Corpus

The test corpus in this paper was collected as part of a larger study (n = 86), which compared the complete Writing Pal system to the AWE component of the system. Here, we focus on the pretest essays produced by these participants. All participants were high-school students recruited from an urban environment located in the southwestern United States. These students were, on average, 16.4 years of age, with a mean reported grade level of 10.5. Of the 45 students, 66.7% were female and 31.1% were male. Students self-reported ethnicity breakdown was 62.2% were Hispanic, 13.3% were Asian, 6.7% were Caucasian, 6.7% were African-American, and 11.1% reported other. All students wrote a timed (25-minute), prompt-based, argumentative essay that resembled what they would see on the SAT. Students were not allowed to proceed until the entire 25 minutes had elapsed. These essays contained an average of 340.84 words (SD = 124.31), ranging from a minimum of 77 words to a maximum of 724 words. Finally, these students completed the same vocabulary knowledge assessment as the students in the previous corpus.

3.4 Vocabulary Knowledge Analysis for the Test Corpus

The two TAALES indices (i.e., *mean age of acquisition log score* and the *normed count for all academic word lists*) were entered as predictors of students' Gates-MacGinitie vocabulary knowledge scores. This regression yielded a significant model, F (2, 83) = 8.521, p < .001, r = .413, R^2 = .170. Only one of the variables was a significant predictor in the regression analysis: mean age of acquisition log score [β =.54, t(2, 83)=3.666, p < .001]. This model suggests that the regression model generated with the primary corpus partially generalized to a new data set. One of the indices accounted for a significant amount of the variance in students' vocabulary knowledge scores. However, this variance was smaller than the variance accounted for in the primary corpus.

4. DISCUSSION

Computer-based writing systems provide students with learning environments in which they can receive writing instruction and engage in deliberate practice [10]. One of the major difficulties that developers of these systems face, however, is the ability to provide instruction and feedback that is *personalized* to individual student users. Developers of these systems often rely on NLP techniques to assess the quality of individual essays; however, it has been relatively unclear whether these NLP techniques can be used to assess relevant individual differences among students.

In the current study, we used NLP techniques to develop stealth assessments of students' vocabulary knowledge. Vocabulary knowledge is an important component of the writing process [5, 19]; thus, our aim was to determine whether we could assess and model individual differences in this knowledge by calculating the lexical sophistication of students' essays. Specifically, an automated text analysis tool was used to analyze the lexical properties of the essays. This tool (TAALES) provided information about the lexical sophistication of the essays at multiple levels (e.g., *word frequency*, *range*, *n-gram frequencies*, *academic language*, and *psycholinguistic word information*). The results revealed that these indices were able to significantly model students' vocabulary knowledge scores. Additionally, these findings were able to predict students' vocabulary scores on a separate data set.

The TAALES correlation analysis revealed that there were 45 lexical sophistication indices that significantly correlated with students' vocabulary knowledge. This is important, because it indicates that individual differences in students' vocabulary knowledge could be detected by analyzing the lexical items that students used in their essays. Further, the regression analyses revealed that the *psycholinguistic word information* and *academic language* indices provided the most predictive power in the model (as opposed to simple measures of word frequency or familiarity), with indices of age of acquisition and academic words accounting for 44% of the variance in the vocabulary scores. Thus, students with greater vocabulary knowledge tended to produce essays with words that are judged to be acquired later in life and were more academic in nature.

Importantly, the follow-up regression analysis revealed that these two TAALES indices accounted for a significant amount of the variance in vocabulary scores for a separate corpus of student essays. In particular, the age of acquisition variable was able to account for approximately 17% of the variance in students' vocabulary knowledge scores. This finding provides confirmation that the automated lexical sophistication indices could be used across two separate data sets to model vocabulary knowledge.

It is important to note, however, that this variable accounted for a significantly smaller amount of the variance in this test corpus than in our primary corpus. This suggests that individual differences may manifest in the properties of students' essays in different ways depending on the specific context. For instance, in this study, the students who produced essays for the two corpora were in college and high school, respectively. Thus, variations in vocabulary knowledge might have influenced the high school and college students' writing process differentially based on the other knowledge, skills or strategies that they had available to them. The results of this follow-up analysis suggest, therefore, that computer-based learning environments may need to rely on separate models for students from different populations. Although the same techniques may be able to be used for all student groups (e.g., the use of NLP), the specific indices in the models may need to be modified across different populations.

Overall, the results from the current study suggest that NLP indices can be utilized to develop stealth assessments of students' skills. When taken together, two indices of lexical sophistication accounted for nearly half of the variance in students' vocabulary knowledge scores. These findings are important, because they indicate that students' individual differences can manifest in the ways that they produce essays. Thus, linguistic analyses of essays (and any other natural language input) may provide useful information about individual students' knowledge and skills. Here, we only analyzed students' vocabulary knowledge at pretest (i.e., before they received any training or feedback). In the future, additional studies will be conducted to specifically examine how these stealth assessments of vocabulary knowledge will change throughout training and how they will serve to inform consistently updating student models.

An additional area for future research lies in the assessment of other individual difference variables. In the current study, we solely analyzed the lexical properties of students' essays because we were focusing on one particular individual difference measure: vocabulary knowledge. In future studies, however, it will be important to consider additional linguistic indices that may be related to other specific constructs of interest. For instance, if we aim to model students' attitudes during writing practice, lexical sophistication indices may provide little valuable information. Instead, we may turn to measures of semantic information, such as the tone or themes found in the essays. Similarly, if we are assessing students' reading comprehension skills, it may be more fruitful to include cohesion indices, which describe the degree to which information in a text is explicitly connected.

In conclusion, the current study utilized the NLP tool, TAALES, to investigate the efficacy of NLP techniques to inform stealth assessments of vocabulary knowledge. Eventually, we expect that this stealth assessment will enhance our student models within the W-Pal system and allow us to provide students with more pointed feedback and instruction. More broadly, the current study suggests that NLP techniques can (and should) be used to help researchers and system developers build stealth assessments and student models in computer-based learning environments. These models can ultimately be used to provide more personalized and adaptive computer-based instruction for students.

While a wealth of studies awaits to answer myriad questions on *how* to construct the most powerful models of individual differences without having to administer the tests, this is a strong step forward in demonstrating the feasibility of such stealth measures.

5. ACKNOWLEDGEMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080589 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.

6. REFERENCES

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

[2] 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.

[3] National Assessment of Educational Progress. 2009. The Nation's Report Card: Writing 2009. Retrieved Nov. 5, 2012, nces.ed.gov/nationsreportcard/writing.

[4] National Assessment of Educational Progress. 2011. The Nation's Report Card: Writing 2011. Retrieved Nov. 5, 2012, nces.ed.gov/nationsreportcard/writing.

[5] 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, (2014) 663-691.

[6] Flower, L. and Hayes, J. 1981. Identifying the organization of writing processes. In L. Gregg and E. Steinberg (Eds.), Cognitive processes in writing. Erlbaum & Associates, Hillsdale, NJ, 3-30.

[7] Allen, L. K., Snow, E.L., and McNamara, D. S. 2014. The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In J. Stamper, S. Pardos, M. Mavrikis, and B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (London, UK, July 4 -7, 2014). Heidelberg, Berlin, Germany: Springer, 304-307.

[8] 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* (2002), 94, 305–315.

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

[10] Allen, L. K., Jacovina, M. E., and McNamara, D. S. in press. Computer-based writing instruction. In C. A. MacArthur, S. Graham, and J. Fitzgerald (Eds.), *Handbook of Writing Research.*

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

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

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

[14] 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, (2015), 35-59.

[15] Warschauer, M., & Ware, P. 2006. Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 10*, (2006), 1-24.

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

[17] 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).* Springer, Heidelberg, Berlin, 269-278.

[18] Roscoe, R. D., Varner, 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 (2014), 39-59.

[19] Crossley, S. A., Allen, L. K., Snow, E. L., & McNamara, D. S. in press. Pssst…textual Features… there is more to automatic essay scoring than just you! In *Proceedings of the 5th International Learning Analytics and Knowledge Conference (LAK)*, Poughkeepsie, NY.

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

[21] 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 Edition)*. Lawrence Erlbaum Associates, Taylor & Francis Group, New York, NY, 311-323.

[22] Shute, V. J., Ventura, M., Bauer, M. I., and Zapata-Rivera, D. 2009. Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. Cody & P. Vorderer (Eds.), Serious games: Mechanisms and effects. Routledge, Mahwah, NJ, 295-321.

[23] 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, and B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining*, (London, UK, July 4 -7, 2014),Springer Berlin Heidelberg, 241-244.

[24] Allen, L. K., Snow, E. L., and McNamara, D. S. in press. Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. In *Proceedings of the 5th International Learning Analytics and Knowledge Conference (LAK)*, Poughkeepsie, NY.

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

[26] Vanlehn, K. 2006. The behavior of tutoring systems. International journal of artificial intelligence in education, 16 (2006), 227-265.

[27] Crossley, S. A., Allen, L. K., Kyle, K., and McNamara, D. S. 2014. Analyzing discourse processing using a simple natural

language processing tool (SiNLP). *Discourse Processes*, 51, 511-534.

[28] Crossley, S. A. 2013. Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching, 46* (2013), 256-271.

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

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

[31] Varner, L. K., Jackson, G. T., Snow, E. L., & McNamara, D. S. (2013). Does size matter? Investigating user input at a larger bandwidth. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), Proceedings of the 26th Annual Flordia Artificial Intelligence Research Society (FLAIRS) Conference (pp. 546-549). Menlo Park, CA: The AAAI Press.

[32] Crossley, S. A., and McNamara, D. S. Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society. (pp. 1236-1231). Austin, TX: Cognitive Science Society.

[33] 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, (2013), 1010-1025.

[34] Kyle, K. and Crossley, S. A. in press. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* (in press).

[35] MacGinitie, W.H., MacGinitie, R.K., Maria, K., and Dreyer, L.G.: Gates-MacGinitie Reading Test (4th ed.). The Riverside Publishing Company, Itasca, 2000.

[36] Crossley, S. A., Cobb, T., and McNamara, D. S. 2013. Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, 41, (2013), 965-981.