A Multi-Dimensional Analysis of Writing Flexibility in an Automated Writing Evaluation System

Laura K. Allen
Mississippi State University
Mississippi State, MS
USA
laura.allen@msstate.edu

Aaron D. Likens
Arizona State University
Tempe, AZ
USA
aaron.likens@asu.edu

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

ABSTRACT

The assessment of writing proficiency generally includes analyses of the specific linguistic and rhetorical features contained in the singular essays produced by students. However, researchers have recently proposed that an individual's ability to flexibly adapt the linguistic properties of their writing might more closely capture writing skill. However, the features of the task, learner, and educational context that influence this flexibility remain largely unknown. The current study extends this research by examining relations between linguistic flexibility, reading comprehension ability, and feedback in the context of an automated writing evaluation system. Students (n = 131) wrote and revised six essays in an automated writing evaluation system and were provided both summative and formative feedback on their writing. Additionally, half of the students had access to a spelling and grammar checker that provided lower-level feedback during the writing period. The results provide evidence for the fact that developing writers demonstrate linguistic flexibility across the essays that they produce. However, analyses also indicate that lower-level feedback (i.e., spelling and grammar feedback) have little to no impact on the properties of students' essays nor on their variability across prompts or drafts. Overall, the current study provides important insights into the role of flexibility in writing skill and develops a strong foundation on which to conduct future research and educational interventions.

CCS CONCEPTS

• **Applied computing**~Computer-managed instruction
• **Applied computing**~Interactive learning environments

LAK'18, March 7–9, 2018, Sydney, NSW, Australia

ACM 978-1-4503-6400-3/18/03…$15.00
https://doi.org/10.1145/3170358.3170404

KEYWORDS

writing, flexibility, natural language processing, feedback, revision

ACM Reference format:

L.K. Allen, A.D. Likens, and D.S. McNamara. 2018. A multi-dimensional analysis of writing flexibility in an automated writing evaluation system. In *LAK'18: International Conference on Learning Analytics and Knowledge, March 7–9, 2018, Sydney, NSW, Australia. ACM, New York, NY, USA*, 10 pages. https://doi.org/10.1145/3170358.3170404

1 INTRODUCTION

Writing is a critically important aspect of our daily lives. From the text message we send in the morning reminding our roommate to turn off the coffee pot, to the emails, reports, and research papers we produce at our jobs, our society is increasingly reliant on writing as a primary mode of communication. Perhaps not surprisingly, this skill is a strong predictor of individuals' success in both the classroom and the workplace [1, 2, 3]. Unfortunately, many individuals struggle to adequately develop the skills needed to produce high-quality texts. In fact, according to the 2011 National Assessment of Educational Progress (NAEP), nearly a quarter (21%) of high school seniors in the U.S. were unable to meet the standards for basic proficiency in academic writing, and only 3% of students performed well enough to be considered advanced writers [4].

Despite its importance, writing has received considerably less attention than other skills in educational and research settings [5-6]. One reason for the relatively small amount of research on the writing process relates to the complexity of the task and, consequently, the difficulty of objectively assessing individuals' performance and skills. An individual's ability to effectively communicate through text can be difficult to measure accurately – due in large part to the high levels of variability in the context, audience, and purpose of the writing task [6-9]. Assumedly, because of this complexity, we know relatively little about the writing process and how it develops over time [10-11].

In the classroom, this complexity can have significant consequences on developing writers, as they are often unaware of, or inaccurate in their understanding of, the criteria necessary to successfullyadlsfkj complete a given assignment [12-15].

Compared to more well-defined domains, such as mathematics, it is often difficult to understand the criteria for high-quality writing and, consequently, it is difficult to engage in the metacognitive strategies needed to understand and implement feedback, as well as to revise negative writing behaviors.

An additional concern is that this complexity has led researchers, educators, and assessment companies to measure writing proficiency in relatively isolated and non-ecological contexts. For example, the assessment of writing proficiency (particularly within standardized tests) commonly revolves around the analysis of the linguistic and rhetorical features of an essay in one particular, relatively non-ecological context – a timed, argumentative essay. This poses a serious problem because research suggests that the characteristics of high-quality writing can (and often do) vary across raters, authors, assignments, and contexts [8-9, 14, 16].

Recently, researchers have proposed that a writer's ability to flexibly adapt might more closely capture their skill [8-9]. In particular, the linguistic flexibility hypothesis has been presented – the idea that skilled writing is related to a flexible use of linguistic style, rather than a static set of specific text properties [9]. To test this hypothesis, the researchers leveraged natural language processing and dynamic modeling to capture variability in students' use of narrative style across multiple essay prompts. The results of their initial study provided support for our hypothesis. Namely, they revealed that individuals' flexible use of linguistic properties across writing assignments was associated with their reading and writing skills, as well as their prior knowledge of the topic.

To build a deeper understanding of the role of flexibility in the writing processes, however, there remain multiple questions to be answered. For instance, along what textual dimensions do individuals naturally vary in their language? Are these dimensions similar or different to those that vary across multiple drafts of the same document? What is the role of feedback in linguistic flexibility? Finally, how does this flexibility across dimensions interact with individuals' literacy skills?

The goal of the current study is to address some of these questions by examining linguistic flexibility across multiple dimensions and time points. In particular, we examine the textual dimensions along which individuals vary on separate essay drafts, as well as how this flexibility relates to students' prior literacy skills. Further, we test whether the dimensions of *between-task flexibility* (i.e., across different essay prompts) are similar or different to those that represent *within-task flexibility* (i.e., across original and revised drafts of an essay). A final aim of this study is to examine the role of lower-level feedback (i.e., spelling and mechanics) on these linguistic features of student essays. Specifically, we examine whether students who are given access to spelling and grammar feedback during the writing process produce texts that differ from their peers along the tested linguistic dimensions. Underlying all aspects of our study is the assumption that better writers will be aware of scaffolds afforded by linguistic text properties at multiple levels and will flexibly exploit these linguistic properties across multiple writing tasks.

Below, we provide a brief overview of automated writing evaluation (AWE) systems, which provide the context for the current study. We then describe the current study and present our results and interpretations in light of prior research.

1.2 Automated Writing Evaluation

Researchers and educators have developed computer-based writing tools, such as automated writing evaluation (AWE) systems, to increase opportunities for students to engage in deliberate writing practice and subsequently to alleviate some of the pressures facing writing instructors due to growing class sizes [17]. These tools have been developed with a variety of goals in mind, ranging from automated test assessment to strategy training [18-20]. For instance, automated essay scoring (AES) systems focus on the automatic scoring of students' essays and are typically used by high-stakes testing companies to score essay components targeted by standardized tests [21-23]. These systems rely on natural language processing (NLP) and machine learning techniques to model the scores that expert human raters would assign to essays based on their structure and content [18; 22-24].

More recently, AES systems have expanded beyond these assessment contexts and have been integrated with educational learning environments, such as AWE systems [23, 25] and intelligent tutoring systems (ITSs) [19]. AWE systems allow students to practice writing essays and receive summative and formative feedback on their individual essays, and ITSs build on these systems by providing individualized instruction and practice. Overall, the primary goal of these computer-based learning environments is to move AES systems beyond summative essay assessments to provide students with increased opportunities for deliberate practice with formative feedback and instruction.

Although a wealth of research has been conducted to validate the accuracy of the scores provided by these AES systems, much less attention has been paid to the pedagogical and rhetorical elements of the AWE and ITS systems that use these scores. In fact, these systems have faced significant criticism, which has often centered around their exclusive focus on analyzing the writing product without much consideration for the communicative context surrounding this text, such as the processes that led to the final essay, the individual differences among the users, and the audience the text is meant to address [21, 26]. These are valid criticisms and point toward avenues for much needed research on the efficacy of computer-based writing systems in learning environments. In particular, if researchers are to accept the criticism that essay tasks should be assessed within particular communicative contexts, then they must also question the validity of their current automated essay scoring methods (i.e., relying on specific linguistic properties to model human scores) and consider more flexible methods of assessing and responding to student writing.

1.3 Writing Pal

An overarching aim of the current research is to improve the validity and adaptivity of the Writing Pal (W-Pal) system. W-Pal

is an ITS that was developed to deliver explicit writing strategy instruction and practice to high school and early college students [19, 27]. Contrary to the majority of computer-based writing systems (see [17] 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).

W-Pal offers strategy instruction that emphasizes the three primary phases of the writing process: prewriting, drafting, and revising. 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). Each of these 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 students can use to improve their writing skills.

After students have viewed the lesson videos, they can unlock mini-games that allow them to practice using these writing strategies in isolation before applying them in the context of a complete essay. 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.

One of the primary features of W-Pal is its AWE component (i.e., the essay practice component). This W-Pal component 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 W-Pal for grading. The W-Pal algorithm [28] then calculates a variety of linguistic indices related to the student's submitted essay and provides both summative and formative feedback that is related to the strategies they have learned.

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.

Formative feedback is an important component of writing development, as it provides important information to writers about components of high-quality writing, as well as actionable recommendations for how to improve writing quality. Examples of these recommendations include: generating ideas and examples, maintaining cohesion, 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 W-Pal has found that this training results in improved essay scores, strategy knowledge, and revising strategies [19, 27, 29].

1.4 Current Study

We examine essay writing in the context of the Writing Pal to develop a deeper understanding of how developing writers flexibly vary the linguistic properties of their essays across drafts as well as assignments (i.e., different essay prompts). Further, we examine whether these properties of their writing vary according to students' literacy skills or the presence of on-line low-level feedback.

In this study, we adopt a multi-methodological approach that relies on NLP techniques to investigate the properties of students' essays across multiple linguistic dimensions. Our approach is to consider the notion that there are multiple linguistic dimensions of the texts that students produce. Some surface-level features relate to the characteristics of the words and sentences in texts and can alter the style of the essay, as well as influence its readability and perceived sophistication. Further, discourse-level features can be calculated that go beyond the words and sentences. These features reflect higher-level aspects of the writing such as the degree of narrativity in the essay.

In the current study, students wrote and revised six essays in the AWE component of W-Pal and were provided with both summative and formative feedback on their writing. Additionally, half of the students had access to a spelling and grammar checker feedback during the writing period. None of the students in this study received explicit strategy training from W-Pal. The overall purpose of this study was to address two primary research questions:

1. Along what dimensions, if any, do developing writers flexibly adapt the style of their writing?
    a. Do these dimensions depend on essay prompt or draft?
    b. Does the availability of spelling and grammar feedback while writing have an influence on these linguistic properties of students' essays?
2. Does the nature of students' linguistic flexibility relate to their literacy skills?

Our first hypothesis is that the developing writers in this study will exhibit flexibility across essay assignments at the discourse levels (e.g., narrativity) of the essays. However, they would predominantly exhibit surface-level flexibility (e.g., word and sentence characteristics) at the draft level. This hypothesis stems from the fact that the student writers will use the feedback provided by the AWE system to improve the sophistication of their writing during the revision period, but not engage in the deeper, semantic revisions that would involve changing their approach to answering a particular question. On the other hand, across writing assignments, we hypothesize that writers will choose to answer specific prompts in different ways, which will lead them to demonstrate flexibility at the discourse-level dimensions of their essays. Importantly, we also hypothesize that the way in which students flexibly adapt to these different essay prompts and drafts will interact with their prior literacy skills, such that more skilled students will demonstrate greater flexibility particularly across the stylistic (discourse-level) dimensions.

Second, we hypothesize that students who have access to spelling and grammar feedback while writing will demonstrate less flexibility overall than their peers without access to this feature. This hypothesis follows from the assumption that writing flexibility is a strategic behavior that relies on an individual's assessment of texts at levels that go beyond the surface level. We hypothesize that providing students access to the spelling and grammar checker will prompt them to place a stronger emphasis on the surface-level features of their writing and lead them to engage less flexibly with the writing task.

2 METHOD

2.1 Participants

The participants (n = 131) in this study were high school students recruited from an urban environment located in the southwestern United States. On average, the students were 16.4 years of age (range 14 to 19). Additionally, 65% were female, 65% were Caucasian, 31% were Hispanic, and 4% reported other ethnicities. There were 11 participants who did not have complete data and were, therefore, dropped from the subsequent analyses. Therefore, the sample size for the models reported below was n = 119.

2.2 Study Procedure

The current study was a three-session experiment that took place over the course of 2-3 weeks for each student. During each session, students wrote and revised two essays in the context of the AWE component of W-Pal. In this component of the system, students had access to a word processor that prompted them to write an essay in response to an SAT-style argumentative essay prompt. For instance, one prompt asked students to develop an argument regarding whether competition or cooperation was more important for success.

All students were given 25 minutes to complete their initial essay draft. They then received automated summative and high-level strategy feedback from the system, and were given an additional 10 minutes to revise their essay. In addition to the high-level feedback, half of the participants received spelling and mechanics feedback during the writing and revising periods, similar to the spelling and grammar feedback provided by the Microsoft Word processor.

2.3 Reading Comprehension Assessment

Students' reading ability was assessed using the Gates-MacGinitie (4th ed.) reading skill test [30]. This 48-item multiple-choice test assessed students' reading comprehension ability by asking students to read short passages and then answering two to six questions about the content of the passage. These questions were designed to measure both shallow and deep level comprehension. All students were given standard instructions, which included two practice questions. This test was a timed task that gave every student 20 minutes to answer as many questions as possible. The Gates-MacGinitie Reading Test is a well-established measure of student reading comprehension, which provides information about students' literacy abilities (α= .85-.92) [31].

2.4 Automated Text Analyses

Coh-Metrix [32] is a computational text analysis tool that was developed, in part, to provide stronger 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 challenges and linguistic scaffolds contained in a given text. Previous work with Coh-Metrix suggests that multiple dimensions coordinate within texts to affect subsequent comprehension performance. To account for these multiple text dimensions, Graesser and colleagues (2011) [33] 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 [34].

2.4.1 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.

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

2.4.3 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.

2.4.4 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.

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

2.5 Statistical Analyses

To address our research questions, we conducted linear mixed-effects models using the lme4 package in R [35]. The purpose of the linear mixed-effects models was to examine the extent to which students varied the linguistic properties of their essays across and within writing tasks (i.e., across separate essay prompts/assignments and between original and revised drafts of their essays). Additionally, students' experimental condition (i.e., the spelling and grammar feedback) served as a fixed effect in our analyses, which allowed us to examine whether having access to the spelling and grammar checker during the writing process influenced the way in which students responded to the different writing tasks along multiple linguistic dimensions.

3 RESULTS

Percentage scores on the reading comprehension test suggest that students varied considerably in their literacy skills, ranging from a minimum score of 10% correct to a maximum score of 100% (M = 57.30, SD = 19.93). To confirm that there were no differences in

reading abilities across the experimental groups, we calculated a between-subjects ANOVA, which revealed that there were no significant differences between the reading scores for the students in the no spelling and feedback condition (M = 59.24, SD = 20.32) and the spelling and feedback condition (M = 55.19, SD = 19.44), F (1, 117) = 1.23, p = 0.27.

3.1 Linguistic Flexibility across Writing Assignments

We assessed the influence of prompt (i.e., essay writing assignment) and experimental condition (i.e., spelling and grammar feedback) on each of the linguistic dimensions of students' six original essay drafts using linear mixed-effects models. As fixed effects, we entered prompt, experimental condition (no spelling/grammar feedback coded as -0.5; spelling/grammar feedback coded as 0.5), and reading ability (grand mean centered reading comprehension scores) into the model. As random effects, we included intercepts for the individual subjects. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. For each of the models listed below, significance was determined using likelihood ratio tests between each model and a reduced model. These models are described below.

For each linguistic dimension, a null model was created, which included random intercepts for each of the participants. Model 2 added the fixed effect of prompt. Model 3 added the fixed effect of reading ability (students' reading comprehension scores). The full model (Model 4) added an interaction term between reading ability and essay prompt to determine whether the effect of prompt on the linguistic dimension depended on students' reading comprehension skills. Two additional models were tested for each of the linguistic dimensions to determine whether there was a main effect of experimental condition or an interaction between condition and prompt. Neither of these models improved model fit and are therefore not presented in the current paper.

The results of the likelihood ratio tests are presented below. For each of these analyses, the first essay that students produced during the study (i.e., an essay in response to a prompt about competition and cooperation) was coded as the reference group. Thus, the fixed effect of prompt examines differences between this prompt and the other prompts that students responded to in the study. Regardless of the chosen reference group, however, the overall model results obtained by the likelihood ratio tests remain the same.

3.1.1 Narrativity. Participants' original essays had an average narrativity score of 77.89 (SD = 19.79) across the six prompts. To assess whether these narrativity scores varied across the prompts, we compared the null model to Model 2, which contained the fixed effect of prompt. Model 2 significantly improved model fit over the null model, χ^2 (5) = 136.495, p < .001, which confirmed that there was a main effect of prompt on the narrativity scores. This suggests that students were varying the style of their essays in response to the different prompts that they were assigned during the study. The addition of the fixed effect of reading ability in Model 3 further improved model fit, χ^2 (1) = 20.850, p < .001 over Model 2, indicating that more skilled readers produced texts that were, on average, less narrative than did less skilled students.

The full model (Model 4) including the interaction between reading ability and prompt only marginally improved model fit over Model 3, χ^2 (5) = 10.087, p = 0.073; however, there was a significant interaction effect between reading ability and two of the prompts. This suggests that, for some of the essay prompts, students' method of adapting their narrative style differed as a function of reading comprehension skill.

3.1.2 Syntactic Simplicity. On average, students produced essays with a syntactic simplicity score of 42.98 (SD = 23.94), indicating that students tended to produce essays with complex syntactic constructions. As with the narrativity analyses, the log likelihood ratio tests between the null model and Model 2 indicated that there was a significant effect of prompt on the syntactic simplicity in students' essays, χ^2 (5) = 70.926, p = < .001. Thus, students did not produce essays with the same form of syntactic constructions for each prompt; rather, they adapted their language across the essay prompts. Model 3 indicated that there was a significant effect of reading ability on the syntactic simplicity in students' essays, χ^2 (1) = 3.964, p < .05; however, as with the narrativity analyses, the addition of the interaction term between reading ability and prompt in Model 4 only marginally improved the fit of the model, χ^2 (5) = 9.904, p = .078. Thus, while reading comprehension skills interacted with students' syntactic flexibility for some of the essay prompts, this interaction effect was not strong enough to significantly improve model fit beyond the previous models that only included the fixed effects of prompt and reading ability.

3.1.3 Word Concreteness. The word concreteness of the essays that students produced was generally low (M = 24.79, SD = 22.22), which suggests that students relied heavily on abstract language in their writing. There was a significant main effect of prompt on the word concreteness in students' essays, χ^2 (5) = 107.907, p < .001, indicating that students were varying the concreteness of the words that they were using across the six essay prompts. However, neither the addition of the main effect of reading ability in Model 3, χ^2 (1) = 3.154, p = .076, nor the interaction between reading ability and prompt, χ^2 (5) = 2.013, p = 0.847, improved the fit over this prompt-only model.

3.1.4 Referential Cohesion. The average referential cohesion score for the essays that students produced was 61.22 (SD = 28.62). Further, there was a significant main effect of prompt on these referential cohesion scores, χ^2 (5) = 115.211, p < .001. This suggests that students varied the referential cohesion in their essays in response to the different prompts that they were assigned. Further, there was a main effect of reading ability on the referential cohesion in these essays, χ^2 (1) = 16.532, p < .001, indicating that more skilled students produced essays that contained less referential cohesion compared to their less skilled peers. However, the interaction in Model 4 did not significantly improve model fit, χ^2 (5) = 6.865, p = 0.231 indicating that students'

differential responses to these prompts did not vary as a function of their reading ability.

3.1.5 Deep Cohesion. On average, students produced essays with high deep cohesion scores (M = 83.54, SD = 20.42). However, the results of the likelihood ratio test between the null model and Model 2 indicated that these scores varied significantly as a function of the prompt, χ^2 (5) = 48.264, p < .001. There was no main effect of reading ability nor was there an interaction between prompt and reading ability.

3.1.6 Preliminary Discussion. The results of the analyses on students' prompt-based flexibility indicate that students demonstrated flexibility at the prompt level across all five of the linguistic dimensions that were tested. In particular, a model that included a fixed effect provided a significantly better fit of our data compared to one that simply accounted for students' linguistic style based on an individual essay. Further, students' scores on a reading comprehension test were significantly related to the amount of narrativity, syntactic simplicity, and referential cohesion included within their essays. In particular, more skilled students tended to produce essays that were less narrative and referentially cohesive but more syntactically simple than their less skilled peers. Further, the reading comprehension scores interacted with some of the prompts along these dimensions, suggesting that students' literacy skills may have played a role in students' flexibility for some prompts, but not for others.

These results partially support our initial hypotheses. We found that students flexibly responded to the six essay prompts along all of the linguistic dimensions that we tested. As predicted, these results suggest that the linguistic properties of student writing vary based on the prompt to which they are responding as well as individual differences in the students' literacy skills. This effect of prompt was more pronounced than we originally predicted, however, in that it was significant across all five of the linguistic dimensions. This suggests that students were capable of flexibly adapting to different prompt demands across both the surface- and deeper-levels of the texts that they produced.

The results also contradicted a number of our initial hypotheses. First, we did not find that the interaction between reading ability and prompt was strong enough to improve model fit over the previous main-effect models. This interaction was significant for some of the prompt comparisons; however, the overall interaction effect was marginal or non-significant for all of the linguistic dimensions. This suggests that the way in which students adapted to the various prompts was not as strongly driven by their linguistic skills as we had hypothesized. Second, the results did not indicate that there was a main effect or interaction with students' experimental condition as we had originally hypothesized. This suggests that the presence of the spelling and grammar feedback during the writing process did not have an influence on students' use of particular linguistic features in their essays.

3.2 Linguistic Flexibility across Original and Revised Essay Drafts

To examine the influence of draft and experimental condition on of the linguistic properties of students' essays, we calculated linear mixed-effects models that modeled students' original and revised essay drafts. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. For each of the models listed below, significance was determined using likelihood ratio tests between each model and a reduced model. These models are described below.

Because of the influence of comprehension scores on the linguistic dimensions in the previous analyses, we entered reading ability as a fixed effect in the null model. Additionally, we included random slopes for the essay prompts and participants to account for the fact that each of the students responded to the prompts in different ways. Model 2 added the main effect of essay draft (i.e., original v. revised draft) and Model 3 examined whether there was an interaction between reading ability and draft. As in the analyses above, two additional models were tested for each of the linguistic dimensions to determine whether there was a main effect of condition or an interaction between condition and draft. None of these models improved model fit and are, therefore, not included in the current paper. The primary results are presented below.

3.2.1 Narrativity. Model 2 significantly improved model fit over the null model for the narrativity dimension, χ^2 (1) = 4.360, p < .05. This indicates that students increased the degree of narrativity in their essays between their original (M = 77.89, SD = 19.79) and revised (M = 78.39, SD = 19.56) drafts. However, this prompt effect did not interact with students' reading abilities, as indicated by the results of the likelihood ratio test between Model 2 and Model 3, χ^2 (1) = 0.311, p = .577.

3.2.2 Syntactic Simplicity. There was not a significant effect of draft on the syntactic simplicity in students' essay drafts, χ^2 (1) = 1.418, p = .234, nor was there an interaction between draft and reading ability, χ^2 (1) = 0.080, p = .777. The results of these analyses suggest that students did not systematically alter the syntactic constructions within their essays across the original (M = 42.98, SD = 23.94) and revised (M = 43.33, SD = 23.93) drafts.

3.2.3 Word Concreteness. There was a main effect of draft on word concreteness, χ^2 (1) = 5.196, p < .05. This model indicates that students decreased the overall concreteness of the words in their essays between the original (M = 24.79, SD = 22.22) and revised (M = 24.02, SD = 21.14) drafts. This effect did not significantly interact with students' reading ability, χ^2 (1) = 2.341, p = .126, suggesting that both more and less skilled students revised these words in similar ways.

3.2.4 Referential Cohesion. Similar to the results of the narrativity and word concreteness analyses, the results revealed that there was a main effect of draft on referential cohesion, χ^2 (1) = 8.085, p < .01. This indicates that, on average, students increased the degree of referential cohesion in their essays across the original (M = 61.22, SD = 28.62) and revised (M = 62.29, SD = 27.89) drafts. This effect of essay draft did not interact with students' reading ability, however, χ^2 (1) = 0.055, p = .815.

3.2.5 Deep Cohesion. Finally, the results of the deep cohesion analyses revealed that students increased the deep cohesion of their essays across the original (M = 83.54, SD = 20.42) and revised (M = 84.24, SD = 19.78) drafts, χ^2 (1) = 5.064, p < .05. However, there was again no interaction between this effect of draft with students' reading ability, χ^2 (1) = 1.944, p = .163.

3.2.6 Discussion. The results of our analyses on essay revisions revealed that students revised their essays along all of the analyzed linguistic dimensions except for syntactic simplicity. In particular, students increased the narrativity, referential cohesion, and deep cohesion in their essays across drafts, whereas they decreased the concreteness of their writing. These effects provide important information about the nature of students' essay revision periods. In particular, students tended to make revisions that would increase the overall readability of their essays at deeper levels of the text (i.e., narrativity, referential cohesion, deep cohesion). However, for the surface-level properties (i.e., word concreteness and syntax), they either made changes that decreased the difficulty (word concreteness) or did not make changes (syntactic simplicity).

Importantly, the results of our analyses further indicated that the nature of students' revisions did not interact with their reading ability. Although reading ability was a significant predictor in all of the models except for syntactic simplicity, students' reading comprehension scores did not significantly interact with essay draft. This suggests that the way in which students chose to revise their essays was not as strongly driven by their literacy skills as we had originally hypothesized.

Finally, as with the previous analyses, the results did not indicate that there was a main effect of students' experimental condition nor an interaction between condition and essay draft on any of the five linguistic dimensions. Therefore, the presence of the spelling and grammar feedback during the writing process did not seem to have an influence on the types of changes that students made during their writing and revising periods.

4 CONCLUSIONS

In this study, we examined the relations between linguistic flexibility, reading comprehension ability, and spelling and grammar feedback in the context of an automated writing evaluation system. In particular, we analyzed student essays along multiple linguistic dimensions to explore the ways in which they flexibly adapted their language across prompts as well as across drafts. We additionally investigated whether this flexibility varied as a result of students' reading abilities or as a function of the presence of spelling and grammar feedback.

The results confirmed the notion that developing writers demonstrate flexibility across the essays that they produce. Indeed, there was a significant effect of prompt on all five of the linguistic dimensions that we analyzed, suggesting that students did not simply produce essays that followed a "template" for good writing, but rather that they adapted their language in response to the demand characteristics of the prompts they were given. Importantly, these results additionally revealed information about similarities and differences between students' flexibility between and within essay prompts. At the revision level, students made changes to their drafts on all dimensions except for syntactic simplicity. This large overlap between our sets of analyses suggest that students were sensitive to the properties of their essays across both surface- and deep levels and produced and revised their texts accordingly.

Although our results suggest that students made revisions across four out of the five linguistic dimensions, it is also important to note that these students made relatively few revisions to the essays overall. In fact, the null model, which included the fixed effect of reading ability and random slopes for participants and prompts, accounted for over 90% of the variance in the data for all five of the linguistic dimensions. This suggests that the majority of the variability in the essays could be accounted for by student-level characteristics, rather than changes that students made across drafts. This result confirms and extends prior research, which has suggested that developing writers often struggle to meaningfully revise their writing across multiple drafts and often will only respond to feedback on their writing at the surface level. Here, we find that students revised essays along multiple dimensions of the text; however, these revisions were relatively minor and did not result in large differences between the original and revised drafts. Importantly, students in this study were not provided with any training from the W-Pal system. Therefore, a question for future research will be whether students benefit differently from these forms of feedback when they have received explicit training.

Our analyses also indicated that providing students with spelling and grammar feedback had no effect on the properties of their essays nor on their variability across prompts or drafts. This suggests that students were not responding to the lower-level feedback when writing and revising their essays; rather, they were adapting their language based on other factors. This is a critical point, given the high level of importance often placed on spelling and grammar feedback in automated writing evaluation systems. Despite researchers' and educators' common assumption that lower-level feedback will lead to improvements in the quality of students' essays, our results suggest that there were no differences in the essays written by the students who received this feedback and those who did not. This finding provides supporting evidence for recent research on writing instruction, which indicates that spelling and grammar instruction and feedback have little to no effect on the quality of students' writing [40-41]. Graham and Perin (2007), for instance, conducted a meta-analysis, which concluded that that spelling and grammar instruction was the only form of writing instruction that did not have a positive effect on students' writing quality [41].

Finally, our results revealed important insights into the role of literacy skill in students' use of specific linguistic properties in their essays, as well as its relation to their flexibility across and within prompts. First, our results revealed that there were no dimensions on which the prompt by reading ability model significant improved model fit over the main-effect model. This was true for both the prompt-level analyses, as well as the draft-

level analyses. For the prompt-level analyses, however, there were three linguistic dimensions (i.e., narrativity, syntactic simplicity, referential cohesion) for which their effects depended on reading ability for some, but not all, of the prompts. This suggests that students' linguistic flexibility across and within prompts (writing assignments) may be driven by a combination of demand characteristics from the prompt (which may presumably impact writers in similar ways), as well as individual differences in students' literacy skills (which may lead writers to produce texts in different ways).

Taken together, the results of our analyses emphasize the importance of examining the writing process from a multi-dimensional and contextualized perspective. Contemporary methods of assessing writing often focus on the analyses of essays in highly de-contextualized scenarios, which place a heavy emphasis on the specific linguistic properties of the essays rather than on students' use of different textual features across varied communicative contexts. In this study, the linguistic properties of students' writing varied as a function of prompt and reading ability. These results call into question the validity of assessing writing proficiency as simply a linear combination of linguistic features. Instead, this study suggests the need for research on the writing process that more carefully considers the nuances that constrain students' behaviors, such as their individual differences, the presumed audience, and the nature of the writing assignment.

Although these results are promising, there are a number of limitations that should be addressed in future research. First, the prompts to which students were asked to respond were relatively similar in their style and demand characteristics. Therefore, the type of flexibility that students were demonstrating might not fully reflect the same form of flexibility that is more commonly observed in real-world writing situations. In future research, we aim to build on this study to address these limitations. In particular, we plan to conduct studies that examine how students adapt their language when they are more explicitly prompted to write for different audiences or for different purposes. We will then examine how fine-grained information about intended writing audiences or contexts can alter the types of revisions that students make to texts. For example, do students alter texts along different dimensions when revising for audiences presumed to have low prior knowledge compared to those with low affect or motivation? These and other similar questions will be the target of future research in this area.

A second limitation of the current research relates to our claims about the degree of flexibility that students demonstrate across the essays and drafts in this study. Because we have not compared these students to other groups (e.g., professional writers, younger students), it is difficult to know how flexibility changes as writing skills develop. It may be the case, for example, that the degree of flexibility that individuals demonstrate significantly increases as they become better writers. Alternatively, however, the possibility remains that writers will reach a threshold for writing flexibility wherein this skill is no longer as important among more skilled writers. These and related questions remain to be answered in future research. These studies will provide a means through which we can better understand the relationship between writing skill and flexibility by understanding how they develop together.

Overall, the work presented in this project provides important insights into the role of flexibility in writing skill. Along with future research, these studies have the potential to enhance our theories of literacy and the roles of context and perspective taking in this process. Our ultimate goal is to leverage this improved understanding of the writing process to develop a stronger foundation for writing research. Results from this type of research can help to advance our understanding of the complexity of writing and discourse and help to inform educational interventions for literacy.

ACKNOWLEDGMENTS

This research was supported in part by the Institute of Education Sciences (R305A120707) and the Office of Naval Research (ONR N000141712300). Any opinions, conclusions, or recommendations expressed are those of the authors and do not necessarily represent views of either IES or ONR. We also thank Cecile Perret, Rod Roscoe and Jianmin Dai for their help with the data collection and developing the ideas found in this paper.

REFERENCES

[1] Gesier, S., & 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.

[2] Light, R. J. (2001). *Making the most of college: Students speaking their minds.* Cambridge: Harvard University Press.

[3] Powell, P. (2009). Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication, 60*, 664-682.[

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

[5] Graham, S., Harris, K. R., & Santangelo, T. (2015). Research-based writing practices and the common core. *The Elementary School Journal, 115*(4), 498-522.

[6] National Commission on Writing. (2004). *Writing: A ticket to work. Or a ticket out.* College Board.

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

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

[9] Allen, L. K., Snow, E. L., & McNamara, D. S. (2016). The narrative waltz: The role of flexibility on writing performance. *Journal of Educational Psychology, 108*, 911-924.

[10] Shanahan, T. (1984). Nature of the reading-writing relation: An exploratory multivariate analysis. *Journal of Educational Psychology, 76*, 466-477.

[11] Shanahan, T. (2016). Relationships between reading and writing development. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research (2nd ed.)* (pp. 194-207) NY; Guilford.

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

[13] Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187-207). New York: Guilford Press.

[14] Varner (Allen), L. K., Roscoe, R. D., & 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.

[15] Wong, B. (1999). Metacognition in writing. In R. Gallimore, L. P. Bernheimer, D. L. MacMillan, D. L. Speech, & S. Vaughn (Eds.), *Developmental perspectives on children with high-incidence disabilities* (pp. 183-198). Mahwah, NJ: Erlbaum.

[16] Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write high quality essays. *Written Communication, 31*, 181-214.

[17] Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of Writing Research*, (2nd ed.), (pp. 316-329). New York: The Guilford Press.

[18] Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*, 3-35.

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

[20] Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. *Handbook of automated essay evaluation: Current applications and new directions*, 36-54.

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

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

[23] Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and future directions*. New York: Routledge.

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

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

[26] 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.

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

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

[29] Allen, L. K., Crossley, S. A., Snow, E. L., & 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.

[30] MacGinitie, W. H., & MacGinitie, R. K. (1989). Gates MacGinitie reading tests. Chicago: Riverside.

[31] Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology, 94*, 3-13.

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

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

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

[35] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48.

[36] Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*, 481-506.

[37] Bridwell, L. (1980). Revising strategies in twelfth grade students'' transactional writing. *Research in the Teaching of English, 14*, 197-222.

[38] Crawford, L., Lloyd, S., Knoth, K. (2008). Analysis of student revisions on a state writing test. *Assessment for Effective Interventions, 33*, 108-119.

[39] Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College Composition and Communication, 31*, 378-388.

[40] Crossley, S. A., Kyle, K., Allen, L. K., & McNamara, D. S. (2014). The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 300-303). London, UK.

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