Reading and Writing
https://doi.org/10.1007/s11145-018-9921-y

Writing flexibility in argumentative essays: a multidimensional analysis

Laura K. Allen[1] · **Aaron D. Likens**[2] · **Danielle S. McNamara**[3]

Abstract

The assessment of argumentative writing generally includes analyses of the specific linguistic and rhetorical features contained in the individual essays produced by students. However, researchers have recently proposed that an individual's ability to flexibly adapt the linguistic properties of their writing may more accurately capture their proficiency. 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 argumentative 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 supposition that skilled writers demonstrate linguistic flexibility across the argumentative 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 argumentative writing skill and develops a strong foundation on which to conduct future research and educational interventions.

Keywords Writing · Flexibility · Dynamics · Linguistics · Natural language processing · Individual differences · Intelligent tutoring systems · Feedback

✉ Laura K. Allen
Laura.Allen@msstate.edu

Extended author information available on the last page of the article

Published online: 06 December 2018

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 work, our society increasing relies on writing as a primary mode of communication. The ability to generate written arguments is a particularly important writing task, as this genre can help individuals persuade, negotiate, debate and resolve conflict between peers and among larger groups (Walton, 1992). Unfortunately, many individuals struggle to develop the skills needed to produce high-quality arguments. According to the 2012 National Assessment of Educational Progress (National Center for Education Statistics, 2012), only 25% of students were considered to be 'proficient' or 'advanced', the level at which students can be expected to meet the expectations of academic writing; further, 21% did not even meet the standards for the 'basic' proficiency category.

Despite its importance, writing has historically received less attention than other skills in educational and research settings (National Commission on Writing, 2004). Recently, however, there has been a sharp increase in research and educational interventions focused on argumentative writing. This is due in large part to students' poor performance on national assessments as well as an increased understanding of the importance of argumentation and literacy skills for workplace proficiency (Biancarosa & Snow, 2006). One of the factors that has typically held back both research and educational interventions related to argumentative writing—and writing more broadly—relates to the complexity of the process and, consequently, the difficulty associated with objectively assessing individuals' performance and skills. The 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 writing tasks (National Commission on Writing, 2004; Allen, Snow, & McNamara, 2014, 2016). Assumedly, because of this complexity, we know relatively little about the writing process, particularly in comparison to other educational domains (Shanahan, 1984, 2016).

Within the context of the classroom, the complexity of argumentative writing can have significant consequences on developing writers, as they are expected to maintain a number of sub goals, such as satisfying the rhetorical demands of the task, addressing the perspectives of the 'other side,' and using appropriate standards to evaluate their own arguments and evidence (Ferretti & Fan, 2016). Consequently, students are often unaware of, or inaccurate in their understanding of, the criteria that are necessary to successfully complete these assignments (Wong, 1999; Roscoe & McNamara, 2013). Given these difficulties, it can be challenging for writers to engage in the metacognitive strategies needed to understand and implement feedback, and to develop the knowledge and skills that are necessary to complete these tasks.

An additional concern is that this complexity has often led researchers, educators, and assessment companies to measure argumentative writing in relatively isolated contexts. For example, the assessment of writing proficiency commonly

revolves around the analysis of the linguistic and rhetorical features of a single argumentative essay (Schoonen, 2005; Kim, Schatschneider, Wanzek, Gatlin, & Otaiba, 2017). In these assessments, students are asked to produce an essay in response to a single prompt; thus, researchers and assessment companies are rarely able to develop robust profiles of students' writing abilities. Instead, they simply capture their ability to respond to a particular prompt in a relatively constrained environment. 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 (Crossley, Weston, McLain-Sullivan, & McNamara, 2011; Varner, Roscoe, & McNamara, 2013; Crossley, Roscoe, & McNamara, 2014; Allen, Jacovina, & McNamara, 2016). Therefore, in order to more adequately capture the components of argumentative writing proficiency, it is important to examine how students adapt to variability across these dimensions.

One of the primary means through which research have examined writing proficiency is to identify the linguistic features of high-quality writing samples (Witte & Faigley, 1981; Deane, 2013; McNamara, Crossley, Roscoe, Allen, & Dai, 2015). This research typically involves the standardized (i.e., rubric-based) scoring of large essay corpora by trained, expert human raters and the subsequent processing of these same essays via natural language processing (NLP) tools. Statistical and machine learning techniques are often used to develop models of the human essay scores based on the NLP indices (Shermis & Burstein, 2003, 2013; Deane, 2013). Results from this line of research have identified a number of linguistic features that commonly characterize high-quality academic writing. For instance, high-quality essays are often associated with a greater number of words, more sophisticated word choices, and fewer spelling and grammar errors than low-quality essays (Haswell, 2000; McNamara et al., 2015).

Importantly, a majority of this research has focused on the argumentative writing genre, presumably because of its prominent role in large-scale testing contexts. However, more recent investigations have begun to examine how these relations differ across genres and contexts (Guo, Crossley, & McNamara, 2013; Biber, Gray, & Staples, 2016; Kim & Crossley, 2018). For instance, Guo et al. (2013) found that essay scores were associated with lexical sophistication in both independent and source-based writing; however, cohesion was only associated with high-quality source-based essays.

Findings from this research have led to a pervasive idea within writing research and educational practice that, because certain linguistic features are often *associated* with essay scores, there must necessarily be a specific way in which high-quality argumentative essays must be written (high cohesion, narrativity, etc.). The influence of this idea can be seen in the content of textbooks, writing manuals, and automated writing evaluation systems, which often place a strong focus on rubrics and feedback that adhere to these strict ideas of 'high-quality writing.' A particularly strong emphasis has been placed on cohesion, for example. Textbooks consistently contain instructions for writers to connect ideas in their essays through the use of explicit word overlap (e.g., the use of similar words and phrases across sentences), as well as through connectives to signal relations amongst ideas.

Linguistic Features and Writing Quality

Despite the general acceptance of these ideas, relations between linguistic features and essay quality can vary widely across raters, authors, assignments, and writing contexts (e.g., Crossley et al., 2011; Crossley, Varner, & McNamara, 2013; Roscoe & McNamara, 2013; Varner et al., 2013; Allen et al., 2014; Crossley, Kyle, Allen, & McNamara, 2014). For example, (Crossley et al., 2014) conducted a cluster analysis showing that there can be multiple linguistic *profiles* of successful (high scoring) essays. Their study provided evidence for the notion that high-quality writing manifests in multiple forms which cannot be defined by a singular set of pre-defined linguistic properties. Researchers have recently extended this idea by arguing that an individual's ability to flexibly adapt the language in their writing might more closely capture their overall level of skill (Allen, Jacovina et al., 2016, Allen, Snow et al., 2016). In particular, the linguistic flexibility hypothesis posits that skilled writing is related to a flexible use of linguistic style, rather than a static set of specific text properties. To test this hypothesis, the Allen and colleagues leveraged natural language processing (NLP) and dynamic modeling techniques to capture variability in students' use of narrative style across multiple essay assignments/prompts. Their findings indicated that individuals' flexible use of narrative properties across argumentative essay tasks was positively associated with reading and writing skills, as well as prior general world knowledge.

Further research is needed to more fully understand the role of flexibility in writing processes, and in turn writing quality. The current study addresses this gap in our understanding by examining linguistic flexibility across multiple dimensions and time points in the context of argumentative essay writing. In particular, we examine the textual dimensions along which individuals vary on separate argumentative essay drafts, as well as how this flexibility relates to students' prior literacy skills. Further, we assess the degree to which 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 students' argumentative essays. Specifically, we examine whether students who are given access to spelling and grammar feedback produce texts that differ from their peers along the targeted linguistic dimensions. Our underlying assumption driving this research is that more proficient writers are aware of scaffolds afforded by linguistic text properties at multiple levels and flexibly exploit these linguistic properties across multiple writing tasks.

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

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 alleviate some of the pressures facing writing instructors due to growing class sizes (Allen, Jacovina et al., 2016). These tools have been developed with a variety of goals in mind, ranging from automated assessment to strategy training (Dikli, 2006; Weigle, 2013; Roscoe, Allen, Weston, Crossley, & McNamara, 2014). Automated essay scoring (AES) systems are typically used by high-stakes testing companies to score essay components targeted by standardized tests (Shermis & Burstein, 2003, 2013; Deane, 2013; Allen et al., 2016; Allen & Perret, 2016). These systems rely on NLP and machine learning techniques to model the scores that expert human raters assign to essays based on their structure and content (Shermis & Burstein, 2003, 2013; Dikli, 2006; Warschauer & Ware, 2006). More recently, AES systems have expanded beyond these contexts and have been integrated with educational learning environments, such as AWE systems (Attali & Burstein, 2006; Shermis & Burstein, 2013) and intelligent tutoring systems (ITSs; Roscoe et al., 2014). 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 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 (Graham, Harris, & Santangelo, 2015; Graham, Hebert, & Harris, 2015).

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 (Perelman, 2012; Deane, 2013). 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.

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 argumentative writing strategy instruction and practice to high school and early college students (Roscoe & McNamara, 2013; Roscoe et al., 2014). Contrary to the majority of computer-based writing systems (see Allen et al., 2016, for a review), W-Pal 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 viewing the lesson videos, mini-games allow students to practice using the targeted writing strategies before applying them in the context of writing an 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 (see "Appendix" for example 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 (McNamara et al., 2015) then calculates a variety of linguistic indices related to the essay and provides both summative and formative feedback 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, by contrast, 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 on how to improve writing quality. Examples of these recommendations include: generating ideas and examples, maintaining cohesion, and employing a variety of different words. The automated formative feedback in W-Pal was specifically developed with this in mind, and provides recommendations that relate to multiple writing strategies (see Fig. 1 for examples of the feedback screen in W-Pal; Roscoe, Varner, Crossley, & McNamara, 2013).

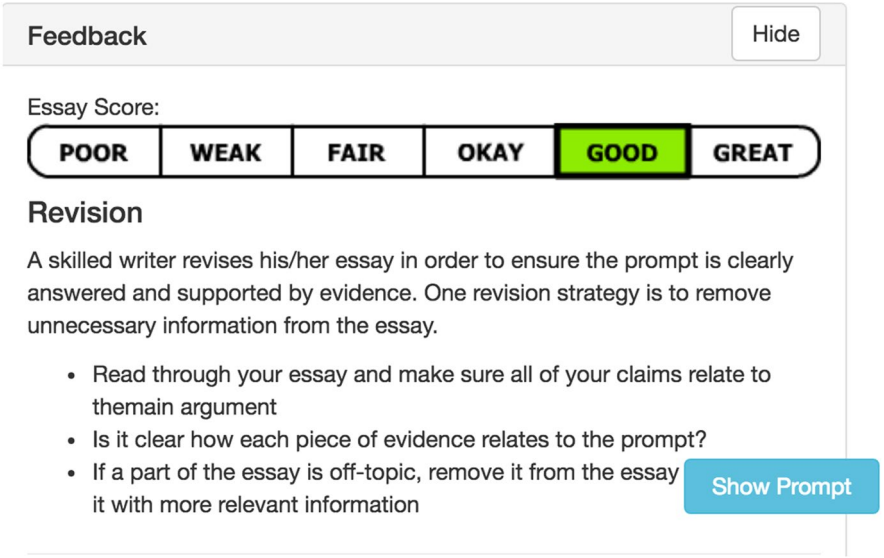

Fig. 1 Screenshot of a Writing Pal feedback report

Table 1 illustrates two examples of feedback that might be provided for (1) an essay that was deemed too short and provides recommendations for strategies to add additional information, and (2) an essay that scored low on the conclusion and needed to work on summarizing the main ideas. Previous research evaluating the efficacy of W-Pal has demonstated that W-Pal training results in improved essay scores, strategy knowledge, and revising strategies (Roscoe & McNamara, 2013; Roscoe et al., 2014; Allen, Crossley, Snow, & McNamara, 2014).

Current Study

We examine essay writing in the context of W-Pal in a study with high school students who are developing their writing skills. Our overarching objective is to develop a deeper understanding of how developing writers flexibly vary the linguistic properties of their argumentative essays across drafts as well as assignments. Further, we examine whether these properties of their writing vary as a function of students' literacy skills and the presence of on-line mechanistic feedback (i.e., grammar, spelling).

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 words and sentences. These features

Table 1 Examples of Writing Pal feedback

Feedback category	Message
Length	Effective writers put forth effort to make sure that the reader can understand the ideas presented. This essay might be expanded in several ways to communicate your ideas more completely One way to expand your essay is to add additional relevant examples and evidence Another way to improve an essay is to provide more details that support your arguments Have you created a flow chart or a writing road map to help you organize your ideas? Trying using the Planning Lesson strategies to make sure your essay is not missing key information
Conclusion	Persuasive essays contain conclusion paragraphs that summarize the main points in the essay. Providing a concluding phrase in the conclusion paragraph signals your reader that your essay is coming to an end Concluding phrases are a great way to begin your conclusion paragraph and to introduce your restated thesis Concluding phrases should clearly tell your reader that your essay is coming to a close Some examples of concluding phrases are: "In conclusion," "In summary," or "As we have seen"

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 argumentative essays in the AWE component of W-Pal and were provided with both summative and formative feedback on their writing. Further, 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 four research questions:

1. Linguistic flexibility across writing assignments:
    a. Along what linguistic dimensions do developing writers flexibly adapt the style of their writing across essay prompts?
    b. Is the nature of students' linguistic flexibility related to their literacy skills?

2. Linguistic flexibility across original and revised essay drafts:
    a. Along what linguistic dimensions do developing writers flexibly adapt the style of their writing across essay drafts?
    b. Is the nature of students' linguistic flexibility related to their literacy skills?

3. Linguistic flexibility and spelling and grammar feedback:

    a. Does the availability of spelling and grammar feedback during writing influence the linguistic properties of students' essays?

4. Linguistic flexibility and essay quality:

    a. Do more *flexible* students produce essays that are judged to be of higher quality than their peers?
    b. Does the relation between linguistic flexibility and essay score depend on the linguistic dimensions being analyzed?

We hypothesize that the nature of students' flexibility will depend on literacy skills and whether they are responding to a new prompt or revising their first draft of the essay. Across writing assignments, we hypothesize that more skilled writers will respond to prompts in different ways, which will be more apparent at discourse-level dimensions (e.g., narrativity, cohesion) of the essays. Specifically, we predict that the ways in which students flexibly adapt to different essay prompts will depend on their prior literacy skills, such that more skilled students will demonstrate greater flexibility, particularly in terms of stylistic (discourse-level) dimensions.

At the draft level, we expect that the students will use the feedback provided by the AWE system to improve the sophistication of their writing during the revision period. However, we expect writing flexibility to depend on both literacy skill and linguistic dimension. We hypothesize that more skilled writers will flexibly revise their essays in terms of both surface-level characteristics (i.e., words, sentences) as well as discourse-level dimensions. By contrast, less skilled students will exhibit flexibility primarily at surface-levels; they are not expected to engage in the deeper, semantic revisions that would require changing their approach to responding to a particular prompt.

Regarding spelling and grammar feedback, 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 with 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. An The alternative competing (null) hypothesis that there will would be no differences between conditions (i.e., with and without grammar and spelling feedback) follows from findings indicating that this form of feedback and instruction has little to no effects on the quality of student writing (Crossley et al., 2014), particularly in the absence of other strategy feedback (Graham & Perin, 2007).

Finally, we hypothesize that students who demonstrate greater linguistic flexibility will produce essays that are judged to be of higher quality. In particular, we anticipate that students who are able to flexibly adapt to the different prompts will demonstrate higher proficiency in writing. Thus, we hypothesize that there will be positive relations between the scores on students' essays and the variability of

their linguistic properties across essays. Further, we hypothesize that this relation between flexibility and essay score will vary based on the linguistic dimensions being assessed. In particular, we hypothesize that flexibility at the lower levels of student essays (i.e., syntactic simplicity, word concreteness) will not be related to essay score; yet, higher level linguistic flexibility (i.e., narrativity, referential cohesion, and deep cohesion) will show significant correlations.

Method

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 14–19) and had just completed 9th (21.6%), 10th (22.4%), 11th (26.1%), or 12th (29.9%) grades. In this participant sample, 65% of the students were female, 65% were Caucasian, 31% were Hispanic, and 4% reported other ethnicities. There were 12 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 $n = 119$.

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 argumentative essays in the context of the AWE component of W-Pal ($M_{length} = 333.159$ words). 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 (see "Appendix" for complete list of argumentative essay prompts). For instance, one prompt (Prompt 1 in "Appendix") asked students to develop an argument regarding whether competition or cooperation was more important for success.

All students were given 25 min to complete their initial essay draft. They then received automated summative and formative strategy feedback from the W-Pal system, and were given an additional 10 min to revise their essay. In addition to the higher 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.

Materials and Measures

Prior Reading Ability

Students' reading ability was assessed using the Gates-MacGinitie (4th ed.) reading skill test (MacGinitie & MacGinitie, 1989). This 48 item multiple-choice test

assessed students' reading comprehension ability by asking students to read short passages and then answer two to six questions about the content of the passage. These questions were designed to measure both shallow and deep level comprehension. Students were given standard instructions, which included two practice questions. This test was a timed task and students were provided students with 20 min to answer as many questions as possible. We calculated students' score as the proportion of questions that were answered correctly on this test. In prior research with this population, we have found that this form of scoring more appropriately captures the important variability in students' scores associated with deep comprehension, rather than the grade level equivalency score. The Gates-MacGinitie Reading Test is a well-established measure of student reading comprehension, which provides information about students' literacy abilities ($\alpha = 0.85$–0.92; Phillips, Norris, Osmond, & Maynard, 2002).

Essay Quality

The quality of students' essays was assessed on a 6-point scale by two independent expert human raters. These raters were college composition instructors with previous experience scoring academic essays and at least 3 years of experience teaching writing who were compensated for their time. The holistic rating scale was developed in order to assess the quality of each essay on a scale from 1 to 6. The raters were given specific instruction on this rubric and given example essays for each score in the rubric (i.e., they were given an example of an essay that had received a score of "1," and another essay that had received a score of "2," etc.). They were instructed that the distance between each score was equal (i.e., a score of 5 is as far above a score of 4 as a score of 3 is above a score of 2).

Raters additionally scored the essays on a number of subscales (ranging from 1 to 6) related to the *introduction*, *body*, *conclusion*, *organization*, *cohesion*, *grammar*, *voice*, *word choice*, and *sentence structure* aspects of the argumentative essays. After receiving instruction on the rubric, the raters practiced using the rubric on a sample set of SAT style essays written on the same prompts as the essays in the current study. The raters were expected to continue with practice until their inter-rater reliability reached a correlation of $r = 0.70$ for holistic scores and all subscales. After the raters had reached an inter-rater reliability of $r = 0.70$, each rater then evaluated the entire set of essays. For the holistic and subscales, all raters had an exact agreement of > 0.9, with the exception of voice which had an exact agreement of 0.84. All raters had 100% adjacent agreement.

Automated Text Analyses

Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) is a computational text analysis tool that was developed, in part, to provide stronger measures of text difficulty (Duran, Bellissens, Taylor, & McNamara, 2007). 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, McNamara, & Kulikowich, 2011) developed the *Coh-Metrix Easability Components* (Graesser et al., 2011). These components provide measures of the principal sources of text difficulty and are well aligned with an existing multilevel framework (Graesser & McNamara, 2011). It is important to note that all of the Easability Component scores range from 0 to 100 with 100 representing the most *readable* and 0 representing more difficult texts.

Narrativity

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

Syntactic Simplicity

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

Word Concreteness

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

Referential Cohesion

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

Deep Cohesion

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

Statistical Analyses

To address our research questions, we conducted linear mixed-effects models using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015). The purpose of these models was to examine the extent to which students varied the

linguistic properties of their essays across and within writing tasks (i.e., across essay prompts/assignments and between original and revised drafts of 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.

For our final set of analyses, we relied on correlation analyses. For these analyses, we aggregated the essay rubric scores across students' essays as well as calculated the coefficient of variation of the linguistic dimensions across the essays. The coefficient of variation, *CV*, is a measure of relative variability and is defined by the following equation, $CV = 100 \times (\sigma/\mu)$, where σ is the standard deviation and μ is the mean (Everitt, 1998). The CV is useful in the current context because it is a unitless number that takes into consideration potential dependencies of a student's variability on their mean behavior (i.e., here, the means of their Coh-Metrix Easability scores). We then conducted Pearson correlations to determine if there were significant relations between essay quality and these linguistic flexibility scores.

Results

Students' essay scores were relatively normally distributed ($M = 3.44$; $SD = 0.74$), reflecting a wide range of possible scores (min = 1.62; max = 5.12). There were no differences in reading comprehension test scores (overall $M = 57.30\%$, $SD = 19.93$; min = 10%; max = 100%) between 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, $p = 0.27$, confirming that the experimental groups were equated in terms of reading skill.

The means and standard deviations for the *Coh-Metrix Easability Components* and essay lengths are presented in Table 2. Additionally, this table reveals the correlations amongst the variables. As can be seen, essay length was not correlated with the Component scores; however, a number of the component scores

Table 2 Descriptive statistics for Coh-Metrix easability components and essay length

Variables	1	2	3	4	5	6	Mean	SD
1. Narrativity	1.00						77.80	19.84
2. Syntactic simplicity	−0.33**	1.00					42.22	24.20
3. Word concreteness	−0.28**	−0.17**	1.00				24.48	22.04
4. Referential cohesion	−0.53**	−0.33**	−0.28*	1.00			62.01	28.46
5. Deep cohesion	0.00	−0.02	−0.05	0.05	1.00		83.76	20.45
6. Essay length	0.00	−0.06	−0.05	−0.03	0.10**	1.00	356.49	139.90

*$p < .05$; **$p < .001$

Table 3 Descriptive statistics for Coh-Metrix easability components across the six prompts

Variables	1	2	3	4	5	6
Narrativity	70.93 (20.88)	69.47 (20.93)	78.34 (18.56)	89.55 (13.96)	80.03 (18.64)	78.69 (18.69)
Syntactic simplicity	46.88 (24.08)	38.06 (23.15)	42.15 (23.69)	34.59 (20.73)	40.89 (24.62)	50.71 (25.53)
Word concreteness	24.35 (20.01)	39.25 (26.42)	14.64 (18.53)	27.71 (22.35)	18.20 (15.73)	22.48 (19.20)
Referential cohesion	61.11 (28.24)	52.69 (31.32)	69.57 (26.33)	75.68 (23.01)	57.60 (28.15)	55.55 (26.46)
Deep cohesion	85.76 (19.57)	83.19 (22.59)	80.61 (21.95)	75.67 (23.81)	92.38 (12.61)	84.91 (16.35)

demonstrated moderate relations. Table 3 provides means and standard deviations for the Easability Components across each of the six prompts.

Linguistic Flexibility Across Writing Assignments

In our first set of analyses, we examined Question 1: Along what linguistic dimensions do developing writers flexibly adapt the style of their writing across essay prompts? Does the nature of students' linguistic flexibility relate to their literacy skills?

We assessed the influence of prompt (i.e., essay assignment)) and literacy skills on each of the linguistic dimensions of students' six original essays 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 skills. Two additional models were tested for each of the linguistic dimensions to determine whether there was a main effect of age or essay length. 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 the analyses, the first essay that students produced (i.e., an essay in response to a prompt about competition) was coded as the reference group. Thus, the fixed effect of prompt examines differences between this prompt and the other prompts to which

students responded. Regardless of the chosen reference group, however, the overall model results obtained by the likelihood ratio tests remain the same.

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 $(5) = 136.495$, $p < 0.001$, which confirmed that there was a main effect of prompt on the narrativity scores. This suggests that students varied 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 $(1) = 20.850$, $p < 0.001$ over Model 2, indicating that more skilled readers produced texts that were, on average, less narrative than did less skilled students ($B = -0.41$).

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.

Syntactic Simplicity

On average, students produced essays with a syntactic simplicity score of 42.98 ($SD = 23.94$), indicating that they tended to produce essays with somewhat 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 the essays, χ^2 $(5) = 70.926$, $p = < 0.001$. 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 $(1) = 3.964$, $p < 0.05$, suggesting that more skilled readers produced more syntactically simple sentences than the less skilled readers ($B = 0.24$). Similar to 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 $(5) = 9.904$, $p = 0.078$. 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.

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. There was a significant main effect of prompt on the word concreteness in students' essays, χ^2 $(5) = 107.907$, $p < 0.001$, indicating that students varied the

concreteness of the words that they used across the essay prompts. However, neither the addition of the main effect of reading ability in Model 3, χ^2 (1) = 3.154, $p = 0.076$, nor the interaction between reading ability and prompt, χ^2 (5) = 2.013, $p = 0.847$, improved the fit over this prompt-only model.

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, $p < 0.001$. 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, $p < 0.001$, indicating that more skilled students produced essays that contained less referential cohesion compared to their less skilled peers ($B = -0.50$). However, the interaction in Model 4 did not 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.

Deep Cohesion

On average, students produced essays with high deep cohesion scores ($M = 83.54$, $SD = 20.42$). 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, $p < 0.001$. However, there was no main effect of reading ability nor was there an interaction between prompt and reading ability.

Preliminary Discussion

Our analyses indicate that students demonstrated flexibility at the prompt level across all five of the tested linguistic dimensions. 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 hypotheses for Question 1. 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 prompt demands across surface and deeper levels of the essays.

However, the results were not fully in line with our hypothesis. Specifically, 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 linguistic dimensions. This suggests (counter to our hypothesis) that the way in which students adapted to the prompts was not as strongly driven by their literacy skills.

Linguistic Flexibility across Original and Revised Essay Drafts

In our second set of analyses, we examined *Question 2*: Along what dimensions, if any, do developing writers flexibly adapt the style of their writing across essay drafts? Does the nature of students' linguistic flexibility relate to their literacy skills?

To examine the influence of draft and condition on 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 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.

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 finding that each of the students responded to the prompts in different ways. Model 2 added the main effect of essay draft (i.e., original vs. revised draft) and Model 3 examined whether there was an interaction between reading ability and draft. The results are presented below.

Narrativity

Model 2 significantly improved model fit over the null model for the narrativity dimension, $\chi^2\ (1) = 4.360$, $p < 0.05$. 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, $\chi^2\ (1) = 0.311$, $p = 0.577$.

Syntactic Simplicity

There was not a significant effect of draft on the syntactic simplicity in students' essay drafts, $\chi^2\ (1) = 1.418$, $p = 0.234$, nor was there an interaction between draft

and reading ability, χ^2 (1) = 0.080, $p = 0.777$. 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.

Word Concreteness

There was a main effect of draft on word concreteness, χ^2 (1) = 5.196, $p < 0.05$. This model indicates that students decreased the 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, $p = 0.126$, suggesting that both more and less skilled students revised these words in similar ways.

Referential Cohesion

Similar to the previous analyses, the results revealed that there was a main effect of draft on referential cohesion, χ^2 (1) = 8.085, $p < 0.01$. This indicates that, on average, students increased the amount 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, $p = 0.815$.

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, $p < 0.05$. However, there was again no interaction between this effect of draft with students' reading ability, χ^2 (1) = 1.944, $p = 0.163$.

Preliminary Discussion

The results of our analyses revealed that students revised their essays along all of the analyzed linguistic dimensions except for syntactic simplicity. 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 revisions. Students tended to make revisions that would increase the overall readability of their essays at deeper levels of the text (narrativity, referential cohesion, deep cohesion). For the surface-level properties (word concreteness and syntax), students either made changes that decreased the difficulty (concreteness) or did not make changes (syntactic

simplicity). In particular, qualitative analyses of students' revisions indicated that students primarily focused on changing the content of their evidence. For instance, in many students' original drafts, they primarily relied on logic as their source of evidence; however, in a number of the revisions, students included either anecdotal stories or facts as evidence to back up their claims. Importantly, these results further indicated that the nature of students' revisions did not interact with reading ability. Although reading ability was a significant predictor in all models except for syntactic simplicity, students' reading scores did not significantly interact with essay draft. This suggests that the ways in which students chose to revise their essays was not as strongly driven by their literacy skills as we hypothesized.

Linguistic Flexibility and Spelling and Grammar Feedback

Our third research question regarded the effects of spelling and grammar feedback on the linguistic properties of students' essays. To address this question, we calculated two sets of linear mixed-effects models to modeled students' essay *assignments* and *drafts*. As fixed effects, we entered experimental condition (no spelling/grammar feedback coded as -0.5; spelling/grammar feedback coded as 0.5). 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 finding that each of the students responded to the prompts in different ways. Models 2a (assignment) and 2b (draft) added the main effect of essay condition (i.e., spelling and grammar feedback) and Models 3a (assignment) and 3b (draft) examined whether there was an interaction between condition and draft.

None of these models containing the fixed effects of condition or the interaction between condition and prompt improved model fit ($p < 0.01$). Thus, the models presented in previous sections containing prompt or draft and reading ability were all significantly better fits to the data than models containing the effect of condition.

Preliminary Discussion

The results of our third set of analyses did not reveal a main effect or interaction with condition on students' essays, supporting the competing (null) hypothesis. This suggests that the presence of the spelling and grammar feedback during the writing process did not influence students' use of particular linguistic features in their essays. Additionally, the results did not reveal a main effect of students' experimental condition nor an interaction between condition and essay draft on students' revisions. Therefore, the presence of the spelling and grammar feedback during the writing process did not appear to influence the types of changes made by students during their drafting and revising phases during writing.

Linguistic Flexibility and Essay Quality

Our final set of analyses addressed our fourth research set of research questions: Do more *flexible* students produce essays that are judged to be of higher quality than their peers, and does the relation between linguistic flexibility and essay score depend on the linguistic dimension being analyzed?

We calculated Pearson correlations between the *CV* (i.e., coefficient of variation) for each of the linguistic dimensions and students' essay scores (holistic and subscale). Results of the correlation analyses suggested that there were significant relations between some of the linguistic flexibility scores and essay scores. Specifically, students' *narrative* flexibility was positively related to the *word choice* subscale scores ($r = 0.187$, $p < 0.05$), and their *referential cohesion* flexibility was positively related to their holistic ($r = 0.204$, $p < 0.05$), grammar ($r = 0.287$, $p < 0.01$), voice ($r = 0.306$, $p < 0.001$), and word choice scores ($r = 0.295$, $p < 0.01$). Conversely, for two of the linguistic dimensions, flexibility and essay scores demonstrated negative relations. *Syntactic* flexibility was negatively related to holistic essay scores ($r = -0.283$, $p < 0.01$), as well as all of the subscale scores (range from $r = 0.260$, $p < 0.01$ to $r = 0.335$, $p < 0.001$). Similarly, *deep cohesion flexibility* was negatively related to holistic essay scores ($r = 0.208$, $p < 0.05$) as well as all of the subscale scores except for *grammar* and *topic cohesion*.

Importantly, all of these correlations remained significant when essay length was controlled for statistically within the analyses, suggesting that the length of the essay was not the factor driving these results. These results suggest that essay quality was positively related to linguistic flexibility along some of the linguistic dimensions (i.e., narrativity, referential cohesion). However, for others (i.e., syntax, deep cohesion), flexibility demonstrated a negative relation to writing quality.

Discussion

In this study, we examined the relations between linguistic flexibility, reading comprehension ability, argumentative essay quality, and spelling and grammar feedback in the context of an automated writing evaluation system. In particular, we analyzed high school student's essays along multiple linguistic dimensions to explore the ways in which students flexibly adapt their language across prompts as well as across essay drafts. We additionally investigated whether this flexibility varied as a function of high school students' reading abilities and the presence of spelling and grammar feedback. Finally, we examined whether students' linguistic flexibility was related to human ratings of the quality of the essays that were produced.

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 presented. Importantly, these results 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. The strong concordance between the various analyses conducted in this study suggest that students were sensitive to the properties of their essays across both surface- and deep levels and consequently produced and revised their texts accordingly.

Although these results suggest that students made revisions across four out of the five linguistic dimensions, it is also important to note that 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 was 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 respond to feedback on their writing only 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 instructional components. 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 high school students with spelling and grammar feedback had no effect on the properties of their essays (as measured by Coh-Metrix) nor on their variability across prompts or drafts. This suggests that this particular sample of students was not responding to the lower-level feedback when writing and revising their essays, at least as measured by Coh-Metrix. Rather, they seemed to be adapting their language based on other factors. For instance, it may be the case that students were responding to the prompts differently based on their own prior experiences in the world or specific factual evidence they were able to recall regarding the prompt topics. Additionally, some students may have had the appropriate metacognitive skills to assess the quality of their original drafts on their own and were therefore able to revise their essays based on this internally-guided feedback. 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 further evidence that spelling and grammar instruction and feedback have little to no effect on the quality of high school students' writing (Graham & Perin, 2007; Crossley et al., 2014). 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. It is important to note, however, that our students are secondary (high school) students; thus, these results may not translate to younger, developing students.

Finally, our results revealed important insights into the role of literacy skill in students' use of 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 significantly 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).

Importantly, the results of our final set of analyses suggest that linguistic flexibility was related to the overall quality of students' essays. This provides supporting evidence for the linguistic flexibility hypothesis recently proposed by Allen and Jacovina et al. (2016). In particular, our results suggest that students' flexibility along certain linguistic dimensions relates to the overall quality of the argumentative essays they produce. Importantly, our results also indicate that this relation between flexibility and essay quality depends on the linguistic dimension that is being analyzed. We found that flexibility along the narrative and referential cohesion dimensions was positively associated with holistic essay scores (for referential cohesion) as well as a proportion of the subscale essay scores. Conversely, flexibility along the syntactic and deep cohesion dimensions was negatively related to holistic and subscale scores. This finding suggests that linguistic flexibility is not a general skill that can refer to all dimensions of an argumentative essay. Rather, the importance of flexibility for argumentative writing skills depends on the specific aspects of the text that are varied by the author. Future research is needed to better understand the contexts in which flexibility is and is not related to the quality of students' writing. Most importantly perhaps, are these relations maintained for other genres of writing?

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 simply as 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. Additionally, because the ordering of the prompts was not randomized in this study, we cannot fully account for the effects of the specific prompts or the effects of linguistic flexibility. 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?

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 not possible to ascertain 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 relations between writing skill and flexibility by understanding how they develop over time.

Finally, a third limitation relates to the amount of time students were given to revise their essays. In this study, students were only given 10 min to revise each of their essays, which is a substantially shorter period of time than would typically be recommended for revision. This choice was made due to limitations on the time that students could be asked to remain in the lab for the study; however, it is possible that this design decision had an impact on the types of revisions in which students engaged. Future research will focus more specifically on the nature of students' revisions in automated writing evaluation systems and will take this limitation into consideration. In particular, in a future study, we plan to have students revise their essays over multiple sessions. Additionally, we will code the nature of these revisions so that we can determine how students revised their essays, and whether this changed over time.

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 (see McNamara & Allen, 2018, for a review). Results from this type of research can help to advance our theoretical understanding of the complexity of writing and discourse (Flower & Hayes, 1981; Hayes, 1996; Kellogg, 2008) and help to inform

educational interventions for literacy (Attali & Burstein, 2006; Roscoe et al., 2014; Shermis & Burstein, 2003, 2013).

Acknowledgements This research was supported in part by IES Grants R305A120707 and R305A180261 as well as the Office of Naval Research (Grant No. N00014-16-1-2611). Opinions, conclusions, or recommendations do not necessarily reflect the view of the Department of Education, IES, or the Office of Naval Research.

Appendix: Essay prompts

General Instructions You will now have 25 min to write an essay on the prompt below.

The essay gives you an opportunity to show how effectively you can develop and express ideas. You should, therefore, take care to develop your point of view, present your ideas logically and clearly, and use language precisely.

Think carefully about the issue presented in the following excerpt and the assignment below.

[Prompt Specific Information]

Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

Prompt 1 While some people promote competition as the only way to achieve success, others emphasize the power of cooperation. Intense rivalry at work or play or engaging in competition involving ideas or skills may indeed drive people either to avoid failure or to achieve important victories. In a complex world, however, cooperation is much more likely to produce significant, lasting accomplishments.

Do people achieve more success by cooperation or by competition?

Prompt 2 All around us appearances are mistaken for reality. Clever advertisements create favorable impressions but say little or nothing about the products they promote. In stores, colorful packages are often better than their contents. In the media, how certain entertainers, politicians, and other public figures appear is sometimes considered more important than their abilities. All too often, what we think we see becomes far more important than what really is.

Do images and impressions have a positive or negative effect on people?

Prompt 3 Loyalty is one of the essential attributes a person must have and must demand of others. Being loyal, faithful, or dedicated to someone or something, is not always easy. People often have conflicting loyalties, and there are no guidelines

that help them decide to what or to whom they should be loyal. Moreover, people may be loyal to something harmful or bad.

Should people always maintain their loyalties, or is it sometimes necessary to switch sides?

Prompt 4 Many people believe that to move up the ladder of success and achievement, they must forget their past, repress it, and let it go. But others have just the opposite view. They see their old memories as a chance to reckon with their past and integrate past and present.

Do personal memories hinder or help people in their effort to learn from their past and succeed in the present?

Prompt 5 When we are young, we learn from parents and teachers that we should wait patiently for what we want. Few people would dispute the wisdom or truth of this teaching. Our society, however, with its mad rush and hurry and its insistence on instant gratification and quick responses, encourages and rewards impatience. Experience teaches us that we should not and do not have to wait.

Is it better for people to act quickly and expect quick responses from others rather than to wait patiently for what they want?

Prompt 6 From talent contests to the Olympics to the Nobel and Pulitzer prizes, we constantly seek to reward those who are "number one." This emphasis on recognizing the winner creates the impression that other competitors, despite working hard and well, have lost. In many cases, however, the difference between the winner and the losers is slight. The wrong person may even be selected as the winner. Awards and prizes merely distract us from valuable qualities possessed by others besides the winners.

Do people place too much emphasis on winning?

References

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.

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, NY: Guilford Press.

Allen, L. K., & Perret, C. A. (2016). Commercialized writing systems. In D. S. McNamara & S. A. Crossley (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 145–162). NY: Taylor & Francis, Routledge.

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.

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

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

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

Biancarosa, G., & Snow, C. E. (2006). *Reading next: A vision for action and research in middle and high school literacy—A report from the Carnegie Corporation of New York* (2nd ed.). Washington, DC: Alliance for Excellent Education.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics, 37,* 639–668.

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.

Crossley, S. A., Weston, J., McLain-Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28,* 282–311.

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

Crossley, S. A., Varner (Allen), L. K., & McNamara, D. S. (2013). Cohesion-based prompt effects in argumentative writing. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), *Proceedings of the 26th annual flordia artificial intelligence research society (FLAIRS) conference* (pp. 202–207). Menlo Park, CA: The AAAI Press.

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

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

Duran, N., Bellissens, C., Taylor, R., & McNamara, D. (2007). In D. S. McNamara & G. Trafton (Eds.), *Qualifying text difficulty with automated indices of cohesion and semantics* (pp. 233–238). Austin, TX: Cognitive Science Society.

Everitt, B. (1998). *The Cambridge dictionary of statistics*. Cambridge, NY: Cambridge University Press.

Ferretti, R., & Fan, Y. (2016). Argumentative writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 301–315). NY: Guilford.

Flower, L. S., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication, 32,* 365–387.

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

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

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *Elementary School Journal, 115,* 524–547.

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools—A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18,* 218–238.

Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication, 17*(3), 307–352.

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

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research, 1,* 1–26.

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37,* 39–56.

Kim, Y. G., Schatschneider, C., Wanzek, J., Gatlin, B., & Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and Writing, 30,* 1287–1310.

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

McNamara, D. S., & Allen, L. K. (2018). Toward an integrated perspective of writing as a discourse process. In M. Schober, A. Britt, & D. N. Rapp (Eds.), *Handbook of discourse processes* (2nd ed.). New York: Routledge.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). Natural language processing in a writing strategy tutoring system: Hierarchical classification approach to automated essay scoring. *Assessing Writing, 23,* 35–59.

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

National Center for Education Statistics. (2012). *The Nation's Report Card: Writing 2011 (NCES 2012-470)*. Washington, DC: Institute for Education Sciences, U.S. Department of Education.

National Commission on Writing. (2004). *Writing: A ticket to work. Or a ticket out*. New York: College Board.

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: Parlor Press.

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.

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.

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.

Roscoe, R. D., Varner, L. K., Crossley, S. A., & McNamara, D. S. (2013). Developing pedagogically-guided threshold algorithms for intelligent automated essay feedback. *International Journal of Learning Technology, 8,* 362–381.

Schoonen, R. (2005). Generaliability of writing scores: An application of structural equation modeling. *Language Testing, 22,* 1–30.

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

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). New York: Guilford.

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

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

Varner, 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.

Walton, D. N. (1992). *Plausible argument in everyday conversation*. Albany, NY: State University of New York Press.

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

Weigle, S. C. (2013). *English as a second language writing and automated essay evaluation. Handbook of automated essay evaluation: Current applications and new directions* (pp. 36–54). New York: Routledge.

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication, 32*(2), 189–204.

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.

Affiliations

Laura K. Allen[1] · Aaron D. Likens[2] · Danielle S. McNamara[3]

Aaron D. Likens
alikens@unomaha.edu

Danielle S. McNamara
dsmcnamara@asu.edu

[1] Psychology Department, Mississippi State University, P.O. Box 6161, Mississippi State, MS 39762, USA

[2] University of Nebraska, Biomechanics Research Building, 6001 Dodge Street, Omaha, NE 68182-0202, USA

[3] Learning Sciences Institute, Psychology Department, Arizona State University, P.O. Box 872111, Tempe, AZ 85287-2111, USA