Examining the Role of Linguistic Flexibility in the

Text Production Process

by

Laura Allen

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2017 by the
Graduate Supervisory Committee:

Danielle McNamara, Chair
Nicholas Duran
Carol Connor
Arthur Glenberg

ARIZONA STATE UNIVERSITY

August 2017

ABSTRACT

A commonly held belief among educators, researchers, and students is that high-quality texts are easier to read than low-quality texts, as they contain more engaging narrative and story-like elements. Interestingly, these assumptions have typically failed to be supported by the writing literature. Research suggests that higher quality writing is typically associated with decreased levels of text narrativity and readability. Although narrative elements may sometimes be associated with high-quality writing, the majority of research suggests that higher quality writing is associated with decreased levels of text narrativity, and measures of readability in general. One potential explanation for this conflicting evidence lies in the situational influence of text elements on writing quality. In other words, it is possible that the frequency of specific linguistic or rhetorical text elements alone is not consistently indicative of essay quality. Rather, these effects may be largely driven by individual differences in students' ability to leverage the benefits of these elements in appropriate contexts. This dissertation presents the hypothesis that writing proficiency is associated with an individual's flexible use of text properties, rather than simply the consistent use of a particular set of properties. Across three experiments, this dissertation relies on a combination of natural language processing and dynamic methodologies to examine the role of linguistic flexibility in the text production process. Overall, the studies included in this dissertation provide important insights into the role of flexibility in writing skill and develop a strong foundation on which to conduct future research and educational interventions.

DEDICATION

I would like to dedicate this dissertation to my family and friends who have supported me throughout this journey. Without your love and encouragement, I would not be the person I am today.

ACKNOWLEDGMENTS

I would first like to thank my advisor Danielle McNamara for all of her support throughout my time in graduate school. You believed in me before I believed in myself and, for that, I cannot thank you enough. I am excited to continue working together and challenging each other to explore new ideas. Next, I would like to thank the other members of my dissertation committee – Drs. Art Glenberg, Carol Connor and Nick Duran. Your varied theoretical perspectives and intense passion for learning have made graduate school so fulfilling and enjoyable. Finally, I would like to thank my fellow students and lab members for all of their support during my time in graduate school. I would like to extend a special thanks to Jen, Russell, Aaron, and Cecile who have supported me, made me laugh, and challenged me to think in new ways. Without you, graduate school would have been far lonelier and far less exciting.

TABLE OF CONTENTS

CHAPTER Page

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 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. Not surprisingly, then, this skill is a strong predictor of individuals' success in both the classroom and the workplace (Geiser and Studley, 2001; Light, 2001; Powell, 2009). 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.

Despite its importance, writing has received considerably less attention than other cognitive skills in both educational and research settings (Graham, Harris, & Santangelo, 2015; National Commission on Writing, 2004). 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. Assumedly, because of this complexity, we know relatively little about the writing process and how it develops over time (Allen, Snow, Crossley, Jackson, & McNamara, 2014; Allen, Snow, & McNamara, 2014; 2016; Shanahan, 1984; 2016).

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 successfully complete a given assignment (Donovan & Smolkin, 2006; Graham, 2006; Varner, Roscoe, & McNamara, 2013; Wong, 1999). 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 difficulty is that this complexity has led researchers, educators, and assessment companies to measure writing proficiency in relatively isolated, non-ecological contexts. For example, the assessment of writing proficiency (particularly in the context of standardized tests) typically revolves around the analysis of the linguistic and rhetorical features of an essay in one particular context – a relatively non-ecological context. This poses a serious problem because research suggests that the characteristics of high-quality writing can (and often do) vary across different raters, authors, assignments, and contexts (e.g., Allen et al., 2016; Crossley, Roscoe, & McNamara, 2014; Varner et al., 2013). Recently, researchers have proposed that a writer's ability to flexibly adapt might more closely capture their skill (Allen et al., 2014; 2016); however, this notion has not been extensively tested.

The goal of this dissertation is to experimentally test the relationship between flexibility and writing skill, as well as to explore the linguistic dimensions on which effective writers flexibly adapt their language. Recent research suggests that the variability of an author' style across multiple prompts is linked to their writing proficiency and higher-level language skills. However, the context of this flexibility is unclear. This dissertation takes an initial step at analyzing this flexibility more closely through multi-dimensional analyses of written text, as well as experimental manipulations of the context surrounding the writing task. Underlying this dissertation project is the assumption that better writers will be aware of the scaffolds afforded by linguistic text properties at multiple levels and will flexibly exploit these linguistic properties across multiple writing tasks. Below I will briefly describe research that has examined the linguistic properties associated with text *readability* and text *quality,* and provide a brief overview of the studies proposed in this dissertation.

The Linguistic Properties of Effective Text-based Communication

A wealth of research has been conducted to examine the linguistic features that contribute to the successful comprehension of texts (see McNamara, Graesser, McCarthy, & Cai, 2014 for a review). Researchers have commonly assumed that these same features will also be related to the quality of written texts. However, this assumption has failed to be supported by the literature (Crossley & McNamara, 2011; McNamara, Crossley, & McCarthy, 2010). For instance, texts that contain more complex sentence constructions have been shown to increase load on working memory, which then results in decreased comprehension (Graesser, Cai, Louwerse, & Daniel, 2006). These same measures of syntactic complexity, however, have also been shown to be related to higher ratings of essay quality by expert human raters (McNamara, Crossley, Roscoe, Allen, & Dai, 2015). This example highlights an important research question in the domains of discourse processing: What can the linguistic features of texts reveal about readers and writers, as well as the relationship between these two skills?

Linguistic Features and Comprehension

Text comprehension is a complex task that has received considerable attention by researchers across a number of research domains from a variety of different perspectives. This process relies on a reader’s knowledge of the language and domain of the text content, but also on the use of skills and strategies that are necessary to leverage this knowledge (Kintsch, 1988, 1998; McNamara, 2004; McNamara & Magliano, 2009). In particular, to develop a deep understanding of a text, a reader must generate connections among the concepts presented in the text, as well as with information that has been activated in long-term memory (i.e., prior knowledge).

The meaning that a reader generates from a text (via these comprehension processes) is commonly referred to as the mental representation – this representation contains: the explicit information provided in the text, the prior knowledge activated during reading, and the inferences generated to connect this information (Kintsch & van Dijk, 1978). The overall *coherence* of this mental representation is positively associated with the degree to which readers activate prior knowledge (from earlier in the text and from the outside world), incorporate this knowledge in their mental representation of the text, and develop connections among the individual propositions (McNamara & Magliano, 2009).

Importantly, skilled and knowledgeable readers are more likely to generate inferences while they are reading (Oakhill & Yuill, 1996), particularly at the global level of the text (Millis, Magliano, & Todaro, 2006). Further, empirical research suggests that linguistic properties of texts can be manipulated to scaffold readers through the text comprehension process (Gernsbacher, 1997; McNamara, Kintsch, Songer, & Kintsch, 1996). McNamara and colleagues (1996) for example, found that increased levels of text *cohesion* helped low prior knowledge readers better comprehend texts, but that decreased levels were beneficial for high prior knowledge readers. It is important to note that *cohesion* and *coherence* are not the same construct (Crossley & McNamara, 2011). As previously mentioned, coherence refers to the connections in the reader's mental representation (Gernsbacher, 1997; McNamara & Magliano, 2009; Zwaan & Radvansky, 1998). Cohesion, on the other hand, refers to the explicit cues in a text that signal readers to make connections among ideas (Halliday & Hasan, 1978). For example, connectives can specify relationships between ideas in a text and provide information about the types of relationships they signify (Longo, 1994).

Beyond this example of text cohesion, researchers have investigated a number of other text features that can influence its readability (Bruner, 1986; Graesser et al., 2006; Graesser & McNamara, 2011; Graesser, McNamara, & Kulikowich, 2011; Haberlandt & Graesser, 1985; Kincaid, Fishburne, Rogers, & Chissom, 1975), such as its syntactic complexity, lexical sophistication, concreteness, and genre. For example, the degree to which a text is narrative or expository has been commonly cited as an important aspect of its readability, with more narrative texts typically being easier to read (Bruner, 1986; Haberlandt & Graesser, 1985). Overall, a wealth of empirical research suggests that the specific properties of texts being read are important and can provide important scaffolds for readers in different situations.

Linguistic Features of High-Quality Writing

Similar to the previously described research on text comprehension, the investigation of linguistic features has played an important role in research on writing (Deane, 2013; McNamara, Crossley, Roscoe, Allen, & Dai, 2015; Witte & Faigley, 1981). In these studies, trained expert human raters typically score large corpora of essays using a standardized rubric. Automated natural language processing (NLP) tools are then used to calculate indices related to the properties of these texts. Finally, statistical and machine learning techniques are used to combine these indices to develop models of the human essay scores (Deane, 2013; Shermis & Burstein, 2003; 2013). Findings from these studies have revealed important information about the linguistic properties of high-quality academic texts. For example, at a basic level, essays that receive higher scores tend to be characterized by a greater number of words, better organization, and fewer spelling and grammar errors than lower scoring essays (Haswell, 2000; McNamara et al., 2015).

Given the findings from the comprehension literature (along with anecdotal and intuitive assumptions), researchers and educators have commonly assumed that high-quality essays would also be characterized by the linguistic features associated with greater *readability* (high cohesion, narrativity, etc.). This assumption has been corroborated by the widespread production of textbooks and writing manuals, which devote large sections of their material to the description of these linguistic scaffolds. Cohesion, in particular, has received considerable attention among writing instructors. Many textbooks detail the need for writers to guide readers through their essay with the use of explicit overlap among sentences (i.e., the use of similar words to avoid confusion by the reader), as well as through frequent use of connectives to signal action and relationships.

Despite the widespread acceptability of these assumptions, empirical evidence has typically not been supportive of them. For example, correlations between expert essay scores and cohesion are typically non-significant or even negative, and component measures of text readability follow similar patterns (e.g., Crossley & McNamara, 2010, 2011). In fact, high-quality essays are often positively correlated with aspects of text difficulty, such as less frequent and concrete words, lower cohesion, and more complex syntactic structures.

Integrating the Comprehension and Production Processes

One often-overlooked difference between the reading and writing processes relates to the *context* of the communication task. Unlike text *comprehension*, the production of high-quality texts requires the reader to consider the context of the assignment, as well as the knowledge and opinions of the particular audience. This demonstrates an important difference between the reading and writing processes, and points toward a potentially important writing skill – namely, flexibility. Indeed, recent research points to the contextual variability of these linguistic features across different audiences, prompts, and assignments (Allen, Snow, & McNamara, 2014; 2016; Crossley et al., 2014). Crossley, Roscoe, and McNamara (2014), for example, found that there were multiple profiles of high-quality writing, which demonstrated different linguistic properties. This evidence points toward the need to examine writing in a more situated context. Although high-quality essays in standardized test context may contain many similar properties, it may be more important to consider a writer's ability to adapt when measuring their writing skills.

In line with this notion, research should consider whether and how writers adapt their writing style according to particular audiences. Research in the comprehension literature suggests that certain text scaffolds are differentially beneficial for audiences based on their knowledge and skill level (McNamara et al., 1996). Thus, an important research question is whether writers are aware of these scaffolds and can leverage them during the writing process. Are skilled readers more aware that connections need to be made in the text and subsequently able to understand when (and for whom) it is appropriate to facilitate these connections in the text? Similarly, are highly knowledgeable students better able to understand when aspects of texts will be more or less difficult to readers? These and other questions remain to be answered.

A final, but important, difference between these two research fields relates to the role of text genre and purpose. Although reading and writing researchers have studied both narrative and expository texts, the role of genre and purpose in these fields (particularly in the field of writing) is often underscored in discussions of empirical results. For example, linguistic scaffolds are often studied in the comprehension literature because the texts are being read to *learn*. However, in other contexts, the role of these linguistic features can vary. For instance, narrative texts typically contain lower explicit cohesive cues because the text is more grounded in familiar concepts and, therefore, easier to follow (McNamara, Graesser, & Louwerse, 2012). Similarly, the linguistic features of high-quality texts will surely vary across different genres, as well as based on the rhetorical strategies taken by the writer. Overall, research suggests that the linguistic features of texts can provide important information about the reading and writing processes. However, more research is needed to develop a better understanding of how these linguistic features vary across different contexts, as well as how flexibility can serve as better measures of writing proficiency.

Overview

This dissertation project is comprised of one published journal manuscript and two additional experimental studies that address the role of flexibility in the text production process. The published journal manuscript in Chapter 2 presents and tests the initial hypothesis that writing skill is associated with students' flexible use of linguistic properties, rather than simply their consistent use of a particular set of linguistic properties. To test this hypothesis, the authors leverage natural language processing and dynamic methodologies to capture variability in students' use of narrative style across multiple essay prompts. The results presented in this chapter provide support for the flexibility hypothesis. In particular, students who were flexible in their use of narrativity across multiple essays also wrote essays of higher quality, whereas inflexible writers tended to write lower-quality essays. Further, more flexible writers performed higher than the more inflexible writers on general assessments of literacy and prior knowledge.

The remaining two chapters of this dissertation build on the study presented in Chapter 2 by examining the writing flexibility hypothesis from multiple perspectives. The purpose of the first study (Chapter 3) is to examine how linguistic flexibility manifests across multiple texts produced by developing writers in an automated writing evaluation (AWE) system. Specifically, the purpose of this study is to answer two primary research questions: 1) Along what dimensions, if any, do developing writers vary their writing style across multiple essays? Does this variability relate to differences in students' comprehension ability? 2) Does the receipt of feedback that prompts students to focus on surface-level (i.e., spelling and grammar) features of their writing have an influence on the nature of this flexibility? In this study, students wrote and revised six essays in an automated writing evaluation (AWE) system designed to provide feedback on student writing. All students received summative and formative (i.e., strategy-based) feedback on their essays before the revision period. Additionally, half of the students had access to a spelling and grammar checker that provides "online" feedback throughout the drafting and revision periods. The purpose of this study is to build upon the previous studies to examine linguistic flexibility across multiple dimensions and in the context of educational settings.

Finally, the study described in Chapter 4 examines how students revise texts for different audiences, as well as whether the properties of these revisions interact with their knowledge of the text content. Participants in this study were provided with two texts – of low and high difficulty – and asked to revise each for two different audiences: a group of university professors or a class of fourth grade students. The aim of this final study is to determine whether students revise the texts in ways that are appropriate for the different audiences. Additionally, this study examines whether students' comprehension skills relate to the types of revisions that they make during the revision period. Overall, the individual studies included in this dissertation project provide important insights into the role of flexibility in writing skill and will develop a strong foundation on which to conduct future research and educational interventions.

CHAPTER 2

THE NARRATIVE WALTZ: THE ROLE OF FLEXIBILITY IN WRITING PROFICIENCY

The study of writing proficiency typically involves the collection of essays that students have written in response to a particular topic, and the subsequent scoring of these essays is based on their linguistic and rhetorical properties. The score that a student receives on this essay is then presumed to serve as a strong proxy for their writing proficiency (Attali & Burstein, 2006). Importantly, however, this essay scoring process is extremely difficult and subjective -- even for trained, expert raters – and therefore may not fully capture the construct of writing proficiency (Huot, 1990, 1996; Meadows & Billington, 2005). Accordingly, an important area of research regards *whether* and *how* writing proficiency can be more reliably captured, particularly emphasizing the specific characteristics of both the individual writers and the texts they produce (Crowhurst, 1990; McNamara, Crossley, & McCarthy, 2010; Rafoth & Rubin, 1984; Witte & Faigley, 1981). Findings from such research can inform our theoretical understanding of the writing process (Flower & Hayes, 1981; Hayes, 1996; Kellogg, 2008; McCutchen, 2000; Swanson & Berninger, 1996), as well as the development and automation of writing quality assessments (Attali & Burstein, 2006; McNamara, Crossley, & Roscoe, 2013; McNamara, Crossley, Roscoe, Allen, & Dai, 2015) and pedagogical interventions for struggling writers (Roscoe, Varner, Crossley, & McNamara, 2013; Shermis & Burstein, 2003, 2013).

One assumption that is commonly held among educators, researchers, and students is that more *proficient* writers produce texts that are easier to comprehend than less proficient writers. This assumption relies on the notion that narrative text properties, such as events, characters, and personal anecdotes, help authors to gain the attention of their readers and, subsequently, make texts more relatable (Newkirk, 1997). Indeed, prior research has confirmed that texts with more narrative elements are typically easier to *comprehend* than informational texts (Bruner, 1986; Graesser, Olde, & Klettke, 2002; Haberlandt & Graesser, 1985). Additionally, the degree to which a text is *narrative* as opposed to *informative* is indicative of its readability across a number of domains and grade levels (Graesser, McNamara, & Kulikowich, 2011). Interestingly, however, the link between narrativity and *essay quality* has failed to be supported by prior literature. Although narrative elements may sometimes be associated with high-quality writing (Crossley, Roscoe, & McNamara, 2014), the majority of research on essay quality suggests that higher quality writing is associated with *decreased* levels of text narrativity, and measures of readability in general (Crossley, Weston, McLain-Sullivan, & McNamara, 2011; McNamara et al., 2013).

One potential explanation for this conflicting evidence lies in the *situational* influence of narrative text elements on writing quality. In other words, it is possible that the frequency of specific linguistic or rhetorical text elements alone is not consistently indicative of essay quality. Rather, these effects may be largely driven by individual differences in students’ ability to leverage the benefits of these elements in the appropriate contexts. In this paper, we hypothesize that writing proficiency is associated with an individual's *flexible* use of text properties, rather than simply the consistent use of a particular set of properties. Some researchers have cited flexibility as a characteristic of strong writers (Graham & Perin, 2007; Graham et al., 2012). Graham and Perin (2007), for instance, claimed "proficient writers can adapt their writing flexibly to the context in which it takes place (p. 9)." However, few studies (if any) have empirically tested this claim. In the current study, we address this research gap by investigating how writing proficiency relates to students' flexible use of *narrativity* across multiple essay prompts.

Writing Proficiency

Writing is a complex and demanding activity that requires individuals to coordinate a number of cognitive skills and knowledge sources through the process of setting goals, solving problems, and strategically managing their memory resources (Flower & Hayes, 1981; Hayes, 1996). Importantly, this writing process differs across individuals. Each student brings different strengths and weaknesses to a given writing task and these variables interact to affect their unique writing processes, as well as the strategies and procedures they utilize to produce effective writing. Individual differences can encompass a broad range of characteristics, from students' degree of prior knowledge (e.g., word and content knowledge, etc.) to their daily and overall affect (e.g., their motivation to succeed). Indeed, many models of writing proficiency attempt to account for the influence of individual differences among students, such as knowledge, skill, and working memory capacity (e.g., Kellogg, 2008; McCutchen, 2000; Swanson & Berninger, 1996).

One important difference between skilled and less skilled writers is their level of reading comprehension skill. Reading and writing are tightly connected cognitive processes (Allen, Snow, Crossley, Jackson & McNamara, 2014; Fitzgerald & Shanahan, 2000; Shanahan & Tierney, 1990; Tierney & Shanahan, 1991); therefore, students who are better at comprehending texts (as well as those who read more frequently) also tend to be better at generating high-quality texts. Similarly, writing proficiency can be influenced by differences in students' vocabulary knowledge (Allen, Snow, Crossley et al., 2014; Graham & Perin, 2007). Students who have access to a greater number of vocabulary words have a greater number of options regarding how they convey ideas.

Strong writers also differ from weak writers in their knowledge of the writing process, including their understanding of writing goals and strategies. For example, Saddler and Graham (2007) found that less skilled writers demonstrated a weaker understanding of writing goals (d = -1.13), were less knowledgeable of the differences between strong and poor writing (d = -.98), and had less knowledge of efficient writing strategies (d = -1.10). Additionally, these less skilled writers wrote lower-quality and shorter essays.

Finally, individual differences in prior world knowledge may influence writing proficiency (McCutchen, 1986; Olinghouse, Graham, & Gillespie, 2015). Olinghouse and colleagues, for instance, recently examined the role of discourse and topic knowledge in the quality and characteristics of 5th grade students' stories, persuasive essays, and informational text. The results of this study suggested that discourse and topic knowledge were important elements of young students' writing skills. Specifically, they found that each of the two forms of knowledge made unique, significant contributions to a prediction of writing quality. These results are important, as they indicate that variability in knowledge can influence the quality of a written text. This is important, particularly in the context of persuasive essay writing, because students who know more about the world can, theoretically, develop stronger arguments, as they have greater access to supporting examples and evidence.

Linguistic Features of High-Quality Writing

Many of these characteristics of skilled writers (e.g., strong reading comprehension skills, etc.) are directly related to their production of specific linguistic properties in essays (Deane, 2013). In particular, more sophisticated linguistic text properties (e.g., cohesion, complex syntax, etc.) are related to higher cognitive functioning. Thus, their presence in an essay is indicative of a student’s ability to more easily produce complex text, which allows them to place a greater focus on higher-level rhetorical and conceptual text properties (Deane, 2013). To this end, many researchers have sought to identify the linguistic properties that relate to high-quality writing (e.g., Applebee, Langer, Jenkins, Mullis, & Foertsch, 1990; Crossley, Roscoe, McNamara, & Graesser, 2011; Ferrari, Bouffard, & Rainville, 1998; McNamara et al., 2010; Varner, Roscoe, & McNamara, 2013; Witte & Faigley, 1981). In these studies, trained, expert human raters typically score essays based on a standardized rubric (e.g., the SAT rubric). The essays are then analyzed for specific linguistic properties, either using computational text analysis tools or human coding. Finally, statistical techniques (e.g., regression analyses, ANOVAs, discriminant function analyses, etc.) are employed to determine whether there are specific linguistic properties that systematically relate to these human judgments of essay quality.

These previous analyses have provided critical information about the linguistic properties of high-quality writing (particularly in the context of academic essays; Applebee et al., 1990; Crossley, Roscoe, et al., 2011; Ferrari et al., 1998; McNamara et al., 2010; Witte & Faigley, 1981). For instance, skilled writers tend to produce longer essays (Crossley, Weston, et al., 2011; Ferrari et al., 1998; Haswell, 2000; McNamara et al., 2010; McNamara et al., 2013) that contain fewer spelling and grammar errors (Ferrari et al., 1998). At the word-level, more proficient writers (i.e., writers that produce higher-quality essays and writers in higher grades) use longer words (Haswell, 2000) that are less frequent and concrete, but are more abstract (Crossley, Weston et al., 2011; McNamara et al., 2010; McNamara et al., 2013). Similarly, previous research has demonstrated that more advanced writers produce essays that contain more complex sentence structures (McCutchen, Covill, Hoyne, & Mildes, 1994). Haswell (2000), for instance, reported that advanced writers produced essays that contained longer sentences and clauses, and McNamara and colleagues (2010) reported that higher-quality essays contained sentences that had a greater number of words before the main verb phrase (i.e., more complex sentence structures).

Finally, specific rhetorical and stylistic text properties have been associated with higher-quality essays. Past studies have found that human ratings of essay quality tend to be negatively related to the frequency of narrative text properties, but positively related to the number of rhetorical structures that focus on contrasted ideas, explicitly stated arguments, conditional structures, and reported speech (Crossley, Weston, et al., 2011; McNamara et al., 2013). Overall, previous research studies reveal that more sophisticated writers (defined by both essay scores and higher grade levels) tend to produce essays that are longer and contain properties that are more indicative of sophisticated lexical, syntactic, and rhetorical choices.

Situational variability of writing quality. Recently, researchers have noted that the text properties associated with essay quality often vary across different raters, authors, assignments, and contexts (e.g., Allen, Snow, & McNamara, 2014; Crossley et al., 2014; Crossley, Weston et al., 2011; Crossley, Varner, & McNamara, 2013; Crossley, Varner, Roscoe, & McNamara, 2013; Varner et al., 2013). Crossley and colleagues (2014), for instance, argued that high-quality essays can take on a number of different forms – in other words, these essays can range quite broadly in their combinations of linguistic properties. To investigate this argument, they employed a cluster analysis approach for the purpose of identifying multiple linguistic *profiles* of successful essays. Their analysis revealed four distinct profiles of successful writers, which were linguistically distinct from one another. They argued that these results provided evidence that successful writing cannot be simply defined by one set of pre-defined linguistic properties -- rather, successful writing can manifest in a number of different ways.

Our hypothesis is that writing proficiency is related (at least in part) to students' *sensitivity* to these different writing styles and, consequently, their ability to flexibly adapt the properties of their essays according to the specific context of the writing task. Writing proficiency, in other words, is partially characterized by an individual's ability to assess the context of their writing task and flexibly call upon various linguistic tools given their knowledge of the constraints and demands of that surrounding environment. For example, if a writer has a strong degree of prior knowledge about the topic for a particular writing assignment, they may not need to employ narrative, story-like properties in order to persuade the reader to take their side on a given argument. On the other hand, if writer is presented with a topic on which they know few explicit facts, they might leverage these narrative story elements for the purpose of engaging their readers and eliciting emotional reactions. Writers in both of these examples could potentially develop successful essays (e.g., they might persuade their readers to take a particular side on an argument); however, the two essays would be composed of vastly different writing styles.

Here, we define writing flexibility as an individual's ability to adapt specific components of their writing in order to craft more effective text. Our argument is that quality texts should not be assessed using a one-size-fits-all formula; rather – successful text communication will depend on a large number of contextual factors, such as the prior knowledge and motivations of the writer and the audience, as well as specific characteristics of the assignment. Importantly, these characteristics of the writing task interact with each other to impact the demands of a particular writing assignment. Thus, writers must assess each writing task on an individual basis to determine the most appropriate strategies and approaches for completing an assignment. In this vein, we argue that more proficient writers will exhibit flexibility in their writing styles across different writing assignments. Our proposal in this paper is that we can measure linguistic flexibility (i.e., the degree to which individuals vary their linguistic style across multiple essays) to serve as a proxy for this broader notion of writing flexibility.

Current Study

The goal of the current study is to test the hypothesis that better writing is associated with increased flexibility of writing style, rather than only a set of static linguistic characteristics. This concept of "flexible" writers is in direct contrast to writers who use a fixed set of linguistic properties within the majority of their essays – in other words, they are *inflexible*. There have been mixed empirical findings regarding the relationship between text *narrativity* (and readability, more broadly) and essay quality. In this study, we suggest that this may be due, in part, to the various demands of the writing assignment. In other words, different writing prompts and assignments may call on different skills and knowledge sources, which can differentially affect the writing strategies and processes engaged by individuals. Thus, we additionally suggest that this flexibility in writing style may result as a function of individual differences related to literacy skills, such as vocabulary knowledge, comprehension ability, and prior world knowledge. Our primary research questions are listed below.

1) How is writing proficiency related to students' flexible use of narrativity?
2) How does this flexible use of narrativity vary as a function of individual differences among students?

We first hypothesize that greater writing proficiency will be positively associated with students' linguistic flexibility across the essays. In particular, we hypothesize that students who vary in their use of narrative language across multiple essays will also produce essays that are rated as higher-quality texts.

Second, we hypothesize that this measure of narrative flexibility will vary as a function of individual differences among the students. This hypothesis follows from the assumption that writing flexibility is a strategic behavior that relates to students' literacy abilities and prior knowledge of a given topic. Thus, students who have developed strong literacy skills will be more likely to assess when it is appropriate to employ specific linguistic and rhetorical devices within individual writing assignments.

This study combines both natural language processing and dynamical techniques to characterize the degree to which students vary in their use of narrativity across 16 timed, argumentative, prompt-based essays. Thus, writing flexibility is measured here in a very specific context. We chose to specifically focus on the narrativity within the essays, because of the previously mixed empirical findings regarding the construct of narrativity in text quality. Crossley and colleagues (2014), for instance, found that one profile of high-quality writing related to a more narrative, story-like, style, whereas a separate profile of essays (of equally high quality) were related to more informative, academic text. Thus, an important research question is whether more proficient writers are able to leverage the benefits of both narrative and informative styles according to the demands of specific writing assignments. For instance, one skilled writer might recognize that she has little fact-based domain knowledge with which to develop evidence on a particular prompt. Therefore, she might construct an essay that relies on personal anecdotes and descriptions that are engaging to her reader. On the other hand, another skilled writer might rely more heavily on fact-based evidence to answer the prompt. In this essay, the writer would use facts to argue a particular perspective on the prompt question. In both scenarios, the resulting essays are high quality and successfully able to argue a particular point to the reader. However, the two writers simply used different strategies to achieve this goal.

An additional note is that this study solely focuses on timed, prompt-based essays. While we argue that this investigation of narrativity is important across a number of different writing genres, we chose to focus our initial analysis on this genre because these essays do not require prior content knowledge of a particular domain. This allows us to more easily tease apart our results in terms of their relationship to writing proficiency, rather than greater knowledge of a particular domain.

Methods of Automated Text Analysis

To address our research questions, we use a combination of natural language processing and dynamic methodologies to examine students' use of narrativity across multiple argumentative essays. Text narrativity is a key component of text readability; therefore, it provides a strong foundation on which to build an understanding of the relations between text readability and essay quality. In this study, we chose to leverage automated text analysis tools to provide a measure of text narrativity. Automated indices provide a quick and reliable alternative to the subjective coding of essays by humans.

Automated measures of text readability and narrativity. In the current study, we employed Coh-Metrix (McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014) to automatically assess the degree to which students' essays were more *narrative* or *informative.* The principal method for automatically measuring text difficulty is the use of standardized “readability” formulas (Hiebert, 2002). These formulas provide a single metric by which the relative syntactic and semantic difficulty of texts can be compared. One of the most common readability formulas is the Flesch-Kincaid Grade Level (FKGL; Kincaid, Fishburne, Rogers, & Chissom, 1975), which calculates word and sentence length to determine text difficulty. This score is a single index that maps onto the grade levels in the U.S. school system. Unidimensional measures, such as FKGL, can simplify the text assignment process by providing teachers a single metric to select grade-appropriate texts for their students.

Despite their simplicity, traditional readability formulas lack the sophistication needed to represent the multiple levels of text difficulty. One problem is that these formulas typically measure the surface-level characteristics of texts, which are solely predictive of students’ superficial text comprehension (i.e., their understanding of the individual words and sentences; Davison, 1984). Most contemporary models of reading comprehension suggest that there are multiple levels of understanding that contribute to the comprehension process (Graesser & McNamara, 2011). However, standard readability formulas often fail to identify the text characteristics that impact students’ understanding at deep levels (e.g., deep cohesion). Further, they provide teachers little guidance on how to diagnose and remediate students’ difficulties. In particular, they give no information on which text properties may be challenging or helpful to individual students.

Coh-Metrix (McNamara & Graesser, 2012; McNamara et al., 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 (McNamara, Graesser, & Louwerse, 2012). To account for these multiple text dimensions, Graesser, McNamara, and Kulikowich (2011) developed the Coh-Metrix Easability Components. These components offer a detailed glance at the primary levels of text difficulty and are well aligned with an existing multilevel framework (Graesser & McNamara, 2011).

Narrativity. The degree of narrativity versus informational content provided within an essay is assessed using the narrativity component score provided by Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011; McNamara, 2013). The narrativity of a text reflects the degree to which a story is being told, using characters, places, events, and other elements that are familiar to readers. This measure is highly related to the use of familiar words, greater world knowledge, and oral language style. Combining many narrative elements within a text can be used to sustain readers' attention by creating uncertainty, excitement, or building suspense (Barab Gresalfi, Dodge, & Ingram-Goble, 2010; Cheong & Young, 2006; Vorderer, Wulff, & Friedrichsen, 1996). Additionally, narrativity allows readers to connect and comprehend action sequences, making it easier to keep track of main characters, plot points, and cause-and-effect relationships (Bruner, 1986; Schank & Abelson, 1995). The degree to which a text is narrative is strongly associated with word familiarity, world knowledge, and oral language.

Because of their engaging and familiar properties, highly narrative texts are considerably easier to read, comprehend, and recall than informative texts (Graesser & McNamara, 2011; Haberlandt & Graesser, 1985). Within the context of essay writing, however, the role of narrativity is less clear. Persuasive essays written with lower degrees of narrativity are typically rated as having higher quality (as judged by expert human raters who use standardized rubrics) than more narrative essays (although not consistently), include more content words (e.g., nouns), and discuss more unfamiliar topics. The use of facts and data as evidence in an essay (as opposed to, for example, personal anecdotes) is associated with more refined rhetorical strategies on the part of the writer, which may serve to explain negative correlations between narrativity and essay scores.

The narrativity component score is calculated in Coh-Metrix based on the results of a previous, large-scale corpus analysis (Graesser, McNamara, & Kulikowich, 2011). In this study, the TASA (Touchstone Applied Science Associates) corpus was used to provide a representative sample of the types of texts that are commonly seen from Kindergarten through 12th grade. This corpus consists of 37,520 texts (average of 288.6 words per text, *SD* = 25.4) that have been classified according to genre and assigned an appropriate grade level. To develop the narrativity score (and the other Easability components), Graesser, McNamara, and Kulikowich (2011) first used Coh-Metrix to analyze the linguistic characteristics of the texts in the TASA corpus (53 measures were used; see Graesser, McNamara, & Kulikowich, 2011 for more specific information about these indices). These indices ranged from basic word level information (e.g., word frequency) to higher-level information about semantic text cohesion. A principal component analysis (PCA) was conducted to reduce these indices to a smaller number of dimensions. The Coh-Metrix measures converged on the PCA with eight principle component scores, accounting for 67.3% of the variability among the texts.

The narrativity Easability Component score consists of 17 Coh-Metrix indices, with loadings ranging from 0.53 to 0.92. These indices provide critical information about the differences between narrative and informational texts. First, narrative texts include more descriptions of actions and events; thus, the narrativity Easability Component assigns its scores (in part) based on the notion that more narrative texts contain more main verbs, adverbs, and intentional events, actions, and particles. Informational texts, on the other hand, are characterized by more unfamiliar content words, often in the form of nouns. An additional characteristic of narrative texts is that they share many characteristics of oral language (Biber, 1988), as evidenced by the increased frequency of familiar words and pronouns in the narrativity Easability Component, as well as the use of simpler sentence constructions.

The resulting narrativity Easability Component score is calculated in the form of a percentile score (ranging from 0% to 100%), with higher scores indicating that the text is more narrative than informative (and likely easier to read) than other texts in the TASA corpus. For instance, a percentile score of 85% means that 85% of the texts in the TASA corpus are likely more difficult than the particular text (at least in terms of its narrativity), and 15% are likely easier to read. Overall, the Coh-Metrix narrativity Easability Component score can serve as a measure of text readability, specifically regarding the degree of story-like elements that are present within an individual text.

Dynamic Analyses

In the current study, we use dynamic systems theory and its associated analysis techniques to analyze the *flexible* relations between the narrative properties of essays and students' writing proficiency. Dynamic methodologies offer researchers a means with which they can characterize patterns that emerge from students' behaviors or interactions (e.g., writing, dialect, or choices) during a learning task. Unlike more traditional statistical measures, dynamic methodologies place a strong emphasis on the role of time in the assessment of behavioral patterns and change. In other words, dynamic analyses focus on the individual fluctuations that occur across time, as opposed to treating behavior as a static (i.e., inflexible) process, as is customary in many traditional statistical approaches (i.e., self-reports). Dynamic methodologies can, therefore, help to contextualize students' behaviors and offer educators and researchers a means of capturing important fine-grained patterns across time.

Although the current study is one of the first to use dynamic analyses to assess *writing flexibility*, these techniques have previously been used across a wide variety of domains as a means to understand the complex patterns that manifest in individuals' behaviors over time (Snow, Allen, Russell, & McNamara, 2014; Snow, Likens, Jackson, & McNamara, 2013; Soller, & Lesgold, 2003; Zhou, 2013). Here, we utilize two dynamic methodologies -- *random walks* and *Euclidian distance*s -- to visualize and classify the extent to which students demonstrate a flexible use of narrative properties across time. Random walks are mathematical tools that are used to *visualize* fine-grained patterns that emerge in categorical data over time (Nelson & Plosser, 1982; Snow et al., 2013). Researchers have used this technique in a variety of domains, such as psychology (Allen, Snow, & McNamara, 2014; Collins & De Luca, 1993), genetics (Lobry, 1996), ecology (Benhamou & Bovet, 1989), and the learning sciences (Snow et al., 2013). For example, geneticists have utilized random walk analyses to investigate how patterns of disease form within gene sequences (Arneodo et al., 1995; Lobry, 1996), and learning scientists have used this methodology to visualize how students' choice patterns within computer-based learning environments vary as a function of their prior skills (Snow et al., 2013).

In order to validate the visualizations offered by these random walk analyses, researchers need to *quantify* these fine-grained patterns of behavior. Euclidian distance analyses offer a metric that is embedded within the random walks that can quantify students' fluctuations as they unfold over time (Allen, Snow, & McNamara, 2014). In this calculation, Euclidian distances for each "step" or movement within a random walk analysis are used to create a distance time series. This time series serves as a quantification for the movements in the categorical patterns visually represented in the random walk.

Method

Participants

The data presented here were collected as part of a larger study (n = 86), which compared the Writing Pal intelligent tutoring system (ITS) to an Automated Writing

Evaluation (AWE) system (Allen, Crossley, Snow, Jacovina, Perret & McNamara, 2015; Allen, Crossley, Snow, & McNamara, 2014; Crossley, Varner, Roscoe, & McNamara, 2013). In this study, we focus on the participants who engaged with the AWE system (n = 45). All participants were high-school students recruited from an urban environment located in the southwestern United States. These students were, on average, 16.4 years of age, with a mean reported grade level of 10.5.

Of the 45 students, 66.7% were female and 31.1% were male. Students self-reported ethnicity breakdown was 62.2% were Hispanic, 13.3% were Asian, 6.7% were Caucasian, 6.7% were African-American, and 11.1% reported "other". All students were recruited from local high schools and publically posted flyers. These students received 10 dollars for their participation in each session of this experiment. Additionally, the students' money was doubled for completing all 10 of the sessions. Thus, the participants in this study each received $200 for their participation.

Study Procedure

The current study was a 10-session experiment that lasted approximately three weeks. During the first session, students completed a pretest that contained measures of writing ability, prior knowledge, reading ability, and literacy skills. Training occurred during the following eight sessions, in which students engaged with the AWE system. During session 10, students completed a posttest, which contained measures similar to the pretest. Previous analyses have indicated that students increased their essay quality, motivation, perceptions of improvement, and self-assessment accuracy across the training sessions (for more thorough information on the results of the training study, see Allen et al., 2015).

Pretest. During session 1, students completed a pretest that lasted approximately one hour in duration and contained a battery of individual difference measures. These measures included demographics, prior knowledge test, writing proficiency (25-minutes SAT-style essay), and literacy skills.

Training. During training (sessions 2-9), students practiced writing 25-minute timed essays on SAT-style prompts. During each of the eight training sessions students wrote and revised two timed essays (i.e., 16 essays). Upon completion of each essay, the AWE system provided students with automated formative feedback. After students examined this feedback they were allotted 10 minutes to revise their essay based on the feedback presented.

Posttest. During session 10, all participants completed a posttest. The posttest comprised measures similar to the pretest, including a writing proficiency test (25-minute SAT-style essay).

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 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; Phillips, Norris, Osmond, & Maynard, 2002).

Vocabulary knowledge. Students' vocabulary knowledge was assessed using the Gates-MacGinitie (4th ed.) vocabulary test (see previous section for reliability; MacGinitie & MacGinitie, 1989). This test includes 45 simple sentences, each with an underlined vocabulary word. Students are asked to read the sentence and choose the word most closely related to the underlined word within the sentence from a list of five choices. All students were given standard instructions, which included two practice questions. This test was a timed task that gave every student 10 minutes to answer as many questions as possible.

Prior knowledge. Students' prior science knowledge was assessed using a 30-item measure of prior knowledge designed for use with high school students. This task has been used previously in work related to reading comprehension and strategy skill acquisition (Roscoe, Crossley, Snow, Varner, & McNamara, 2014). The 30-item multiple-choice measure assesses students' knowledge in the areas of science, literature, and history. The test shows high reliability, with α ranging from .72 to .81. The measure is a modified version of a knowledge assessment used in several studies and validated with over 4000 high school and college students (McNamara, O'Reilly, Best, & Ozuru, 2006; O'Reilly, Best, & McNamara, 2004; O'Reilly & McNamara, 2007; O'Reilly, Taylor, & McNamara, 2006). This version of the assessment was developed in prior work by including items with moderate difficulty (i.e., 30-60% of students could answer correctly) that were correlated with individual difference measures (e.g., reading skill) and performance on comprehension tests. Additional items were obtained from high school textbooks. In this process, 55 multiple-choice questions (i.e., 18 science, 18 history, and 19 literature) were piloted with 15 undergraduates to test item performance. Thirty questions (10 per domain) were selected such that no items selected exhibited either a ceiling (> .90) or floor effect (< .25, chance level). Examples are provided in Table 1.

Table 1.

Examples of questions and answers in prior knowledge assessment

Domain	Question and Answer Choices
Science	The poisons produced by some bacteria are called… a) antibiotics, b) toxins, c) pathogens, d) oncogenes
History	A painter who was also knowledgeable about mathematics, geology, music, and engineering was… a) Michelangelo, b) Cellini, c) Titian, d) da Vinci
Literature	Which of the following is the setting used in "The Great Gatsby"… a) New York, b) Boston, c) New Orleans, d) Paris

Pretest and posttest essay quality. Students writing proficiency was assessed at both pretest and posttest through the use of timed (25-minute) and counterbalanced SAT-style essays (the two essay prompts can be found in Appendix A). The pretest and posttest essays were assessed on a 6-point scale by two independent expert human raters. These raters had previous experience scoring academic essays and were compensated for their time. Additionally, they were college composition instructors with at least three years of experience teaching writing. The holistic rating scale was developed in order to assess the quality of each essay on a scale from 1 to 6 (see http://sat.collegeboard.org/scores/sat-essay-scoring-guide for a copy of the SAT rubric). 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.). Additionally, they were told 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). 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 $r = .70$. After the raters had reached an inter-rater reliability of $r = .70$, each rater then evaluated the entire set of essays. Thus, each essay received two essay scores. Once these ratings were collected, differences between the raters' scores were calculated. All score differences between the raters were less than 2 (i.e., the raters demonstrated an 100% adjacent agreement with the final set). Thus, holistic scores for pretest and posttest essays were calculated by averaging the scores between raters. For the final set, the raters demonstrated a 57% exact accuracy and a 100% adjacent accuracy. Additionally, the raters' final essay scores were significantly correlated ($r = .55$, $p < .001$).

Table 2.

Writing Pal essay prompt order

Session	Essay Prompts
Session 2	**Planning:** Does every individual have an obligation to think seriously about important matters? **Originality:** Can people ever be truly original?
Session 3	**Winning:** Do people place too much emphasis on winning? **Loyalty:** Should people always maintain their loyalties, or is it sometimes necessary to switch sides?
Session 4	**Patience:** Is it better for people to act quickly and expect quick responses from others rather than to wait patiently for what they want? **Memories:** Do personal memories hinder or help people in their effort to learn from their past and succeed in the present?
Session 5	**Heroes:** Should we admire heroes but not celebrities? **Choices:** Does having a large number of options to choose from increase or decrease satisfaction with the choices people make?
Session 6	**Perfection:** Do people put too much importance on getting every detail right on a project or task? **Optimism:** Is it better for people to be realistic or optimistic?
Session 7	**Uniformity:** Is it more valuable for people to fit in than to be unique and different? **Problems:** Should individuals or the government be responsible for solving problems that affect our communities and the nation in general?
Session 8	**Beliefs:** Are widely held views often wrong, or are such views more likely to be correct? **Happiness:** Are people more likely to be happy if they focus on their personal goals or on the happiness of others?
Session 9	**Fame:** Are people motivated to achieve by personal satisfaction rather than by money or fame? **Honesty:** Do circumstances determine whether or not we should tell the truth?

Training essay performance. Training performance in this study was defined as students' average essay score across the 16 essays that were composed in the AWE system. All of the essays that students wrote in this AWE system were timed, SAT-style essays, with prompts that were similar to those given at pretest and posttest (for a list of the prompt topics and the order they were assigned, see Table 2). To score these essays, we used a previously developed algorithm to assign holistic writing scores to these written essays. The algorithm uses variables from Coh-Metrix, the Writing Assessment Tool (WAT), and Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007) to assign essay scores on a scale from 1 to 6. These indices range from word-level properties of the essays, such as the number of infinitives, to higher-level properties, such as the semantic similarity of the paragraphs within the essay. The algorithm was developed using correlation and discriminate function analyses to categorize 1243 student essays that had been previously scored by expert human raters. The resulting models reported exact matches between the human scores and the predicted essay scores with 55% accuracy. Additionally, the models reported 92% accuracy for adjacent matches (see McNamara et al., 2015, for a more thorough description of the algorithm used in this study).

Assessment of narrative flexibility. We used random walk analyses to investigate the flexibility of students' use of narrativity across time. Random walk analyses are mathematical tools that are used to provide visual representations of patterns in categorical data as they manifest across time (Benhamou & Bovet, 1989; Lobry, 1996; Nelson & Plosser, 1982; Snow et al., 2013). In the current study, we first used Coh-Metrix to compute a *narrativity percentile score* (range from 0 to 100) for each essay. We then used this narrativity percentile score to classify each essay into four orthogonal categories (see Table 3). This classification was organized based on the degree of narrativity present in each essay (using the percentile score provided by Coh-Metrix). Each orthogonal category was then assigned to a vector that fell along a basic scatter plot.

Therefore, if an essay received a narrativity score below 25%, this essay was assigned to the vector (-1, 0), whereas an essay that received a score that was greater than 75% narrative was assigned to the vector (0, -1). Once each essay had been assigned to a vector, we calculated a random walk for each student that began at the origin of the scatter plot (0, 0). For each subsequent essay that a student wrote, the walk would "step" in the direction that was consistent with the assigned vector. The resulting walk would represent each student's use of narrativity across the 16 training essays.

Table 3.

Narrativity classification and vector assignment

Essay Narrativity Level	Axis Direction Assignment
Less than 25% Narrativity	-1 on X-axis (move left)
Between 25% and 50% Narrativity	+1 on Y-axis (move up)
Between 50% and 75% Narrativity	+1 on X-Axis (move right)
Greater than 75% Narrativity	-1 on Y-axis (move down)

Figure 1 provides an example of what a random walk might looks like for a student who wrote four training essays. All walk sequences begin at the origin of the scatterplot (see # 0 in Figure 1). The first essay written by the student was low in narrativity (i.e., narrativity percentile score < 25%); thus, the walk takes a step left along the X-axis (see # 1 in Figure 1). The second essay written by the student received a narrativity percentile score between 25% and 50%; this means that the walk takes a step up along the Y-axis (see # 2 in Figure 1). The student wrote a third essay that had a narrativity percentile score between 50% and 75% narrativity. Therefore, the walk takes a step to the right along the X-axis (see # 3 in Figure 1). The fourth and final essay written by the student received a narrativity percentile score between 25% and 50%, which again makes the walk step up along the Y-axis (see # 4 in Figure 1). These rules were used to generate a unique random walk for each of the 45 students, which represented the fluctuations in their use of narrativity across the 16 essays that were written in the AWE system.

Figures 2 and 3 illustrate two random walks that were generated using two students' actual training essays from the current study. These walks represent students' degree of "narrative flexibility" across the training essays.

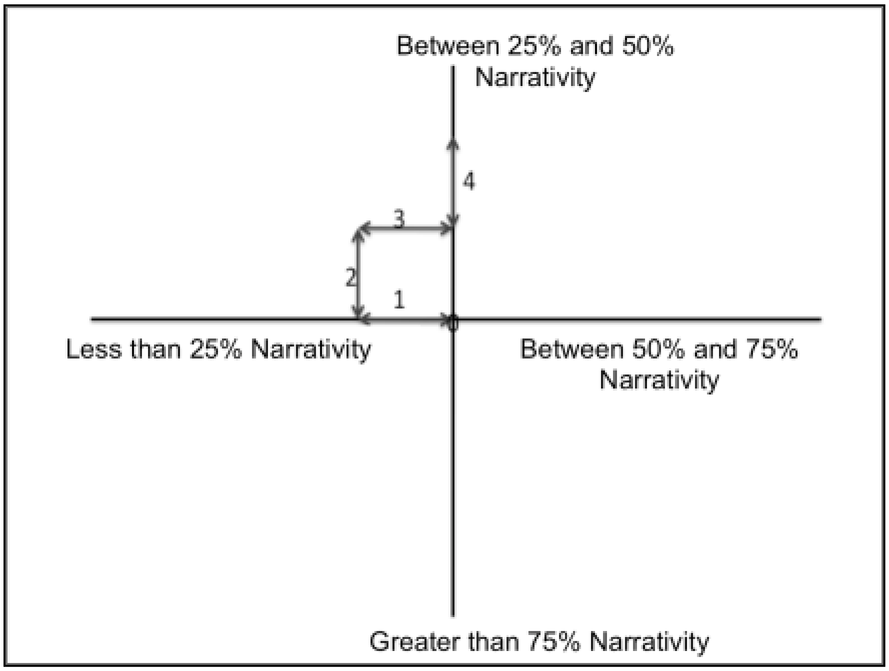
Between 25% and 50% Narrativity
4
3
2
1
Less than 25% Narrativity
Between 50% and 75% Narrativity
Greater than 75% Narrativity

Figure 1. Example random walk

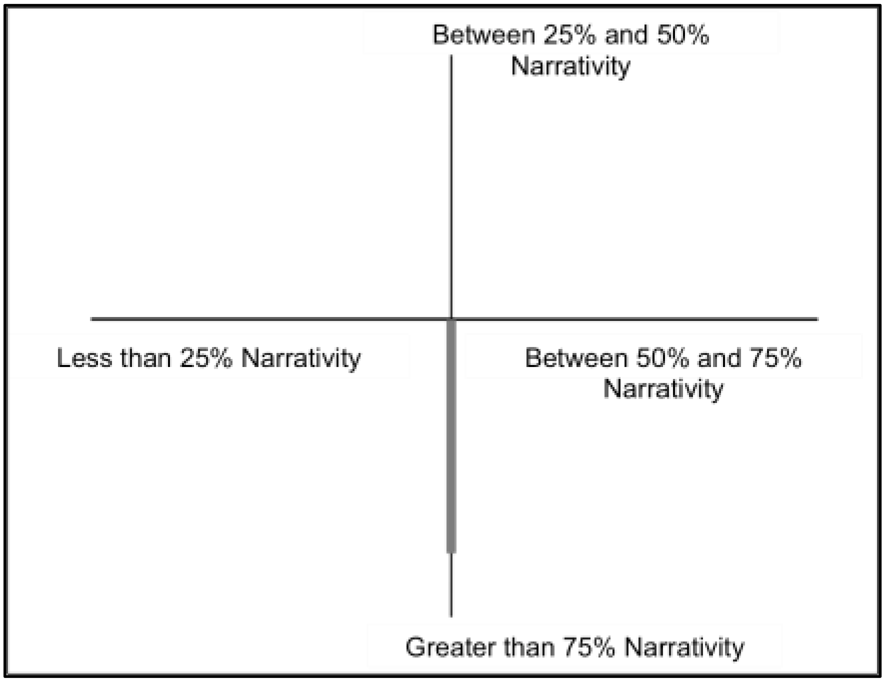
Between 25% and 50% Narrativity
Less than 25% Narrativity
Between 50% and 75% Narrativity
Greater than 75% Narrativity

Figure 2. Example random walk (inflexible narrativity)

Figure 2 illustrates the walk of a student who wrote highly narrative (above 75 narrativity percentile score) essays across each of the training essay assignments. In other words, regardless of the writing prompt, this student employed the same range of narrativity throughout all of her essays. On the other hand, the walk depicted in Figure 3 comes from a student who was highly flexible in the use of narrativity across the 16 essays. As the various factors varied from essay to essay (e.g., the essay prompt), this student employed varying degrees of narrativity to develop arguments and ideas.

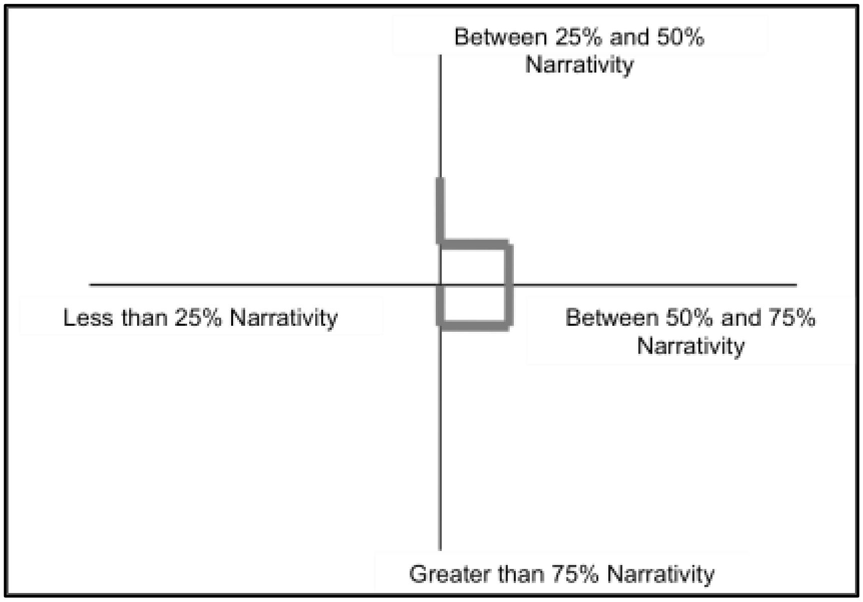
Between 25% and 50% Narrativity
Less than 25% Narrativity
Between 50% and 75% Narrativity
Greater than 75% Narrativity

Figure 3. Example random walk (flexible narrativity)

Euclidian distance measure. The random walks described above provide *visualizations* of the fluctuations in students' use of narrativity across time. To *quantify* these changes in students' writing patterns, distance time series were calculated for each student using Euclidian distance measures. This measure calculated the distances between the origin of the scatter plot (0, 0) and each step in the walk (see Equation 1 below). In this equation, *y* represents the current position of the particle (the end point of the walk) on the Y-axis, *x* represents the particle's position on the Y-axis *and i* represents the *i*th "step" in the walk.

$$\text{Distance} = \sqrt{(y_i - y_0)^2 + (x_i - x_0)^2} \qquad (1)$$

After calculating the Euclidian distance of the steps in each walk, an average Euclidian distance score was calculated for each student's entire walk. Broadly, this measures how far each student "walked" from the origin of the scatter plot across the 16 essays. This resulting distance measure (i.e., a student's *narrative distance score*) was used to represent students' flexibility in their use of narrativity. If a student, for example, employed the same degree of narrativity across all 16 training essays, that student would travel further from the origin, resulting in a high narrativity distance score (see Figure 2 for a visualization of this type of student). Conversely, if a student varied considerably in the use of narrativity across the essays, the resulting narrative distance score would be lower as the fluctuations would cause the walk to remain closer to the origin (see Figure 3 for a visualization of this type of student). Overall, students' distance scores provide information about whether they are varied in their writing style (i.e., lower distance scores and more flexible) or whether they tend to remain inflexible (i.e., consistent) across multiple essays (i.e., higher distance scores and inflexible). It is important to note, that the directionality of students' random walks does not matter as the Euclidian distance measure captures how far (in any direction) students' walks move away from the center point.

The random walk and Euclidian distance analyses used in the current study afford researchers the ability to capture flexibility that would otherwise be missed by traditional (i.e., static) metrics. In particular, random walk analyses capture movements as they take place across time. In this sense, we can analogize the narrative flexibility examined in this study to the dancing of the Waltz. In the Waltz, dancers make multiple movements that result in rotations of the dancers around the floor. Importantly, in the Waltz, skilled dancers do not travel across the room in a straight line. Although this would result in more *efficient* travel, these dancers recognize that in order to perform the dance in the most graceful way, they must make small rotations that result in larger movements across the floor. Additionally, they must make adjustments to their behaviors based on their partner's behaviors, as well as the behaviors of the other dancers on the floor. Thus, in the Waltz, the fine-grained steps and patterns of the dancers are important to its overall aesthetics and success. Similarly, we propose that skilled writers will demonstrate more *flexible* patterns of narrativity across their essays. Thus, rather than consistently producing essays of the same style, these writers will flexibly adapt their behaviors to the demands of the prompt (e.g., based on their own prior knowledge, the audience, etc.). Related to the random walk analyses, if a student generates essays that vary in their degree of narrativity, the student's random walk will hover around the center point of the X, Y axis and contain more movements that change directions. In contrast, a student who is less flexible and consistently generates essays with similar levels of narrativity will demonstrate a random walk that moves in one direction and covers a greater distance along the X or Y axis.

Statistical Analyses

To assess the degree to which writing quality is associated with students' flexible use of narrativity, we calculated random walks, Euclidian distances, Pearson correlations, and regression analyses. The random walk analyses allowed us to *visualize* students' use of narrativity across their 16 essays. Additionally, this random walk allowed us to calculate a Euclidian distance measure, which reveals students' consistency in their use of narrativity across their 16 essays. Pearson correlations were used to assess the relation between flexibility (as defined by the Euclidian distance measure) and essay quality, as well as individual differences in students' prior global knowledge, prior vocabulary knowledge, and prior reading comprehension ability (see Table 4 for descriptive statistics on these pretest and posttest materials). Finally, regression analyses were conducted to follow-up the correlation analyses in order to provide an indication of the variables that accounted for the most variability in the dependent variables (i.e., essay quality and flexibility).

Table 4.

Descriptive statistics for pretest and posttest materials

Measure	**Minimum**	**Maximum**	**Mean (SD)**
Pretest Essay Score	2.00	4.00	2.80 (0.57)
Posttest Essay Score	2.00	4.50	3.10 (0.64)
Reading Comprehension*	21.00	75.00	47.55
Vocabulary Knowledge*	13.00	89.00	56.44
Prior Knowledge (Overall) *	27.00	77.00	51.70
Science Prior Knowledge*	20.00	90.00	52.67
History Prior Knowledge*	10.00	100.00	54.00
Literature Prior Knowledge*	10.00	70.00	48.44

**Score is based on percentage correct*

Results

Random Walks

To visualize and categorize how students varied the narrativity in their writing style, random walk analyses were calculated using the rules described in the previous section (see Table 3) for each student. These walks produced distance measures for each student, which is indicative of how flexible or inflexible the student's use of narrativity was across all 16 essays. Overall these narrative distance measures suggested that students varied considerably in their narrative flexibility, ranging from a minimum narrative distance score of 2.03 to a maximum narrative distance score of 8.50 (*M* = 6.11, *SD* = 1.73). The narrative distance score for each student in this study is plotted in Figure 4 to provide a visualization of the degree to which students varied in their flexible use of narrativity across the 16 training essays.

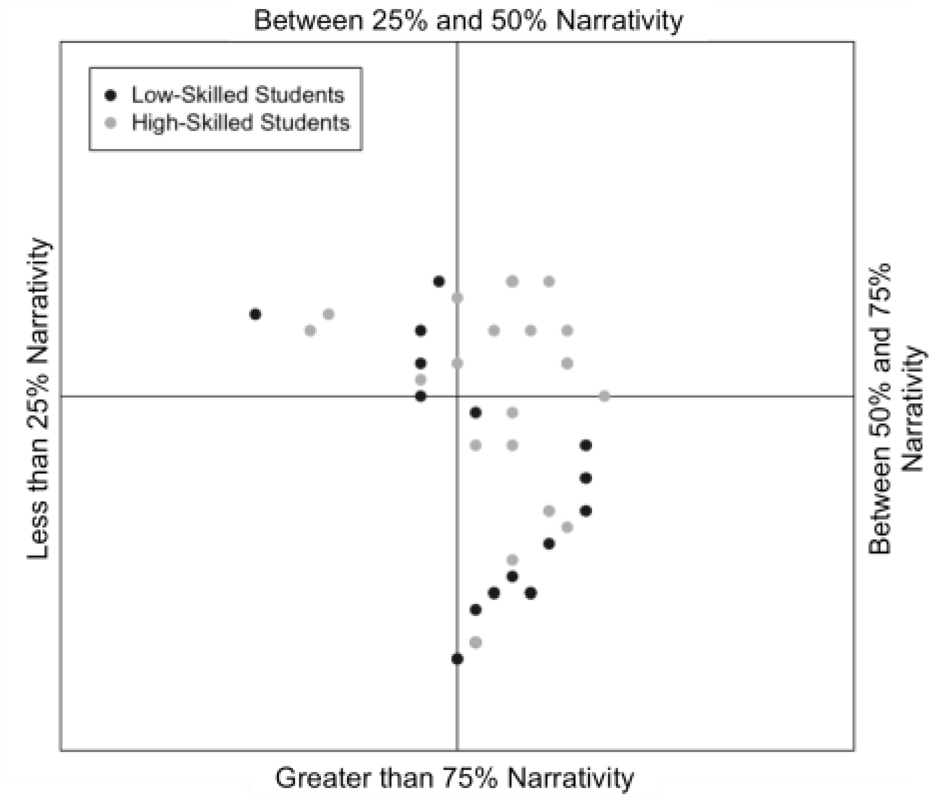

Figure 4. Visualization of students' random walks end points

This variation in narrative flexibility was examined according to students' writing proficiency. To provide a coarse visualization of the flexibility differences between the less and more skilled writers, we created a visualization that compared the narrative distance scores for two groups of students (based on a median split on students' pretest essay scores): *less skilled writers* and *more skilled writers.* To confirm that the visualization was depicting two separate groups of students, a between-subjects ANOVA investigated the difference between these less skilled and more skilled writing ability students' narrative distance scores and revealed that more skilled writers had significantly lower narrative distance scores (*M* = 5.29, *SD* = 1.47) compared to less skilled writers (*M* = 7.02, *SD* = 1.60), *F* (1, 42) = 14.06, *p* = .001, *d* = 1.13.

Figure 4 provides an illustration of these differences between less and more skilled writers. In this figure, less skilled writers are represented as black dots and more skilled writers are represented by light-gray dots. As shown in this image, the less skilled writers (black dots) traveled further from the origin of the scatter plot (0, 0) than the more skilled writers (light-gray dots), who seem to cluster more frequently near the origin. This visualization indicates that the more skilled writers were also the students who were more varied in their use of narrativity across the training essays (i.e., they hovered more around the origin), whereas the less skilled writers travelled much further from the origin and were less flexible in their use of narrativity.

Writing Proficiency

Although the visualization analyses provided preliminary evidence that less and more skilled writers differed in their narrative flexibility, this analysis was based on a median split and, therefore, has potential statistical weaknesses. Median splits pose problems to statistical validity because they create a false dichotomous variable from a continuous variable. Therefore, we conducted further analyses to provide more statistically valid tests of our research questions. Specifically, Pearson correlations were calculated to further assess the validity of these analyses (i.e., to assess the degree to which students' flexible use of narrativity was related to their writing proficiency). We calculated the correlations between students' narrative distance scores and their pretest and posttest essay scores (assessed by the expert human raters), as well as their average scores across the 16 training essays (assessed by the AWE algorithm). Results from these analyses indicated that narrative distance scores were significantly negatively related to the quality of pretest essay scores (r = -.45, p = .002) and training essay scores (r = -.47, p = .001). Overall, these results reveal that skilled writers were more flexible in their use of narrativity across the training essays (i.e., they exhibited lower narrative distance scores). However, the relation between narrative flexibility and essay scores was no longer present at posttest (p = .08). These findings suggest that over the course of persistent writing practice, the relation between flexibility in writing style and essay quality is reduced.

We conducted a stepwise regression analysis with the significant variables as predictors to determine which writing proficiency measures were the most predictive of narrative flexibility, as well as to assess the amount of variance accounted for by these assessments. This analysis yielded a significant model [F (1, 42) = 11.66, p = .001; R^2 = .22] with one variable retained in the final analysis: Training Essay Scores [β = -.47, t (1, 42) = -3.41, p = .001]. Results of this analysis suggested that students' flexible use of narrativity was most strongly predicted by the quality of the essays that they wrote across the eight days of writing practice. Thus, students who consistently demonstrated strong writing proficiency were more flexible in their use of narrativity throughout essay writing practice.

Individual Differences

To further investigate the role of narrativity flexibility in the writing process, we examined its relationship with individual differences known to relate to writing proficiency. Specifically, we calculated Pearson correlations and regression analyses between narrative distance scores and students' pretest scores on assessments of prior world knowledge, vocabulary knowledge, and reading comprehension ability. Results of the correlation analyses suggested that the narrative distance scores were significantly related to all of the pretest measures except for prior knowledge in history and literature (see Table 5). These results suggest that narrative flexibility is related to other literacy skills and knowledge sources, rather than solely related to writing proficiency, as it is strongly associated with performance on assessments of prior science knowledge, as well as literacy skills.

Table 5.

Correlations between distance scores and individual differences

Individual Difference Measure	*r*
Reading Comprehension	-.59**
Vocabulary Knowledge	-.41*
Prior Knowledge (Overall)	-.39*
Science Prior Knowledge	-.44*
History Prior Knowledge	-.27
Literature Prior Knowledge	-.20

p < .05*, *p* < .01**

We conducted a stepwise regression analysis with the significant variables as predictors to determine which individual difference measures were the most predictive of narrative flexibility, as well as to assess the amount of variance accounted for by these assessments. This analysis yielded a significant model [F (1, 43) = 22.47, p < .001; R^2 = .34] with one variable retained in the final analysis: Reading Comprehension [β = -.59, t (1, 43) = -4.74, p < .001]. Results of this analysis suggested that students' flexible use of narrativity was most strongly predicted by ability to read and comprehend texts. Thus, students who entered the writing task with more strategies and knowledge about how to comprehend texts may have had a simpler time adapting their writing styles to various prompts, as they were potentially more aware of the processes engaged by their readers, and thus more strategic in their actions (McNamara, 2013).

Conclusion

Evidence from the field of writing research largely supports the notion that the linguistic properties of texts are generally indicative of the holistic quality of those texts. Indeed, results from a number of studies have pointed toward specific characteristics that predict human judgments of writing quality (Crossley, Roscoe, & McNamara, 2013; McNamara et al., 2010; Witte & Faigley, 1981). The accuracy of these results, however, often varies along with various factors associated with the writing assignment, such as the individual rater or the writing prompt (Crossley et al., 2013; Crossley, Allen, & McNamara, 2014; Varner et al., 2013). In this study, we empirically examined these assumptions through a computational linguistic analysis of students' essays. We leveraged both natural language processing and dynamic methodologies to capture variability in students' use of narrative style and to relate that variability to individual differences in writing proficiency, as well as prior science knowledge and reading comprehension skills.

The results from the current study support our hypotheses that writing proficiency can be characterized (at least in part) by students' flexibility across multiple essay prompts. Namely, students who are more flexible in their use of narrativity tend to receive higher scores on their essays, whereas less flexible writers tend to produce lower-quality essays. Using random walk analyses, we were able to visualize students' flexible or inflexible use or narrativity across the 16 training essays. These analyses revealed the differential patterns exhibited by the less and more skilled writers, with the skilled writers remaining near the origin of the scatter plot and the less skilled writers straying further from the origin. To quantify the findings from this random walk analysis, Euclidian distance measures were calculated. The resulting narrativity distance scores provided confirmatory empirical support for the random walk analyses. In particular, the results demonstrated that less skilled students tended to be more consistent (i.e., inflexible) in the degree to which they used narrative properties (i.e., higher narrative distance scores), whereas more skilled students demonstrated more flexibility in their use of narrativity across the 16 essays (i.e., lower narrative distance scores).

Importantly, the relationship between flexibility and narrativity was no longer apparent at posttest. Our interpretation of this result is that the quality of the students' essays had substantially improved by the time they wrote the posttest essay and, therefore, the individual differences in flexibility were no longer a factor in their posttest essay quality. In other words, the feedback generated by the AWE system was effective. Results from a previous analysis of the larger study (i.e., the comparison between the Writing Pal ITS condition and the AWE condition; Allen, Crossley et al., 2014, 2015;

Crossley et al., 2013; Roscoe & McNamara, 2013) revealed that students' essay scores substantially improved across the training sessions (Allen, Crossley et al., 2015). Additionally, the accuracy of the students' self-assessments of essay quality (compared to the W-Pal algorithm) increased in accuracy over time. This is important, because it potentially indicates that, with practice and feedback, students can become more aware of the quality and specific characteristics of their own writing and therefore produce essays that more effectively address the prompt question.

Additionally, results from the current study revealed important information about individual differences associated with students' flexible use of narrativity. In particular, flexible writers outperformed the inflexible writers on more general assessments of literacy and prior knowledge. Reading comprehension skills were most strongly linked to this flexibility, accounting for 34% of the variance in students' narrative distance scores. This finding suggests that students who were more skilled at comprehending texts and potentially more aware of readers' strategies and cognitive processes (e.g., O'Reilly & McNamara, 2007) were also more easily able to adapt their writing style to match certain contexts.

The results from this study are important for writing researchers and educators, as they indicate that the link between textual properties and writing quality may fluctuate according to the context of a given writing assignment. Accordingly, writing proficiency not only relates to the sophistication of the words and sentences a student produces in a given essay – but is also intimately related to the writer's ability to adapt style, narrative language, and other rhetorical content to individual writing assignments and different audiences. These results may be explained, in part, by the fact that narrativity tends to be an easier writing style to employ for high school students. Thus, when they are faced with multiple difficult writing assignments, they may resort to this easier writing style as a default. Additionally, the results of the individual difference analyses suggest that this flexibility is not exclusively related to writing proficiency; rather, high school students who are more skilled and knowledgeable are better able to adapt the style of their writing according to situational variations.

Although this ability to flexibly adapt to various contexts has been anecdotally cited as an important component of writing proficiency (Graham & Perin, 2007), to date, little to no research has been conducted to empirically test this assumption. The scarcity of research on this topic may be due in large part to the difficulties associated with assessing writing flexibility. First, it requires a longitudinal data set such as the one presented here wherein students are asked to compose multiple essays over time and in response to different prompts. To our knowledge, other such data sets have not been reported in the literature. Second, flexibility is a complex construct to measure. This is particularly true for ill-defined domains, such as writing, which rely on human subjectivity to render judgments about quality and style. Standardized writing assessments typically only measure high school students' writing ability in one particular context and, therefore, cannot be sensitive to fluctuations in style, or in an individual's adaptation to different contexts. If researchers and educators aim to develop assessments that can truly capture students' writing proficiency, it is important to remain sensitive to their ability to adapt their style and language choices according to different assignments and contexts.

The findings and methodologies presented here have important implications for the assessment of students' writing proficiency. In particular, our study indicates that the linguistic properties that interact to predict writing quality may be inconsistent from assessment to assessment. Unfortunately, in their current state, standardized assessments of writing proficiency typically only collect a single writing sample from students. Thus, they are unable to take the construct of writing flexibility into account when making judgments about proficiency. This may constitute a critical oversight. Standardized assessments of writing have a strong influence on students' ability to enter college, as well as their receipt of scholarships and other such opportunities. This study suggests that standardized test developers should aim to develop more sophisticated assessments that can capture students' writing skills across a number of different contexts. Additionally, in the future, the techniques used in the current study may be integrated into a number of educational environments to better assess and improve students' writing skills. For instance, ITSs are computer-based educational environments that provide adaptive instruction and feedback to students based on their skills and performance. Writing-based ITSs might take advantage of this technique to provide feedback that not only looks at students' individual essays, but also captures their flexibility across multiple time points (Allen, Jacovina, & McNamara, 2016).

Notably, the results reported here call for replications across different populations and skill levels of writers and different writing genres. To our knowledge, there are currently no other data sets that would support replications of the current work. Thus, one goal of our future research will be to develop a corpus that contains multiple essays from different genres written by students from varying populations and skill levels. The achievement of this goal will help us to investigate a number of unanswered questions and concerns. Successful authors of persuasive essays, for example, may flexibility adapt their narrativity; however, in other genres, this flexibility may not be a positive writing characteristic. Future research will aim to answer this question, as well as a number of other questions that currently remain unanswered. For example, is it the case that flexibility for *all* linguistic properties is positively related to essay quality? Or, are certain properties more consistently important across a number of different assignments? Further, this study points to the importance of feedback in promoting writing flexibility. This finding prompts the questions: can students be trained to be more flexible in their writing style? What is the role of *feedback* in the promotion of increased writing flexibility? Finally, what cognitive processes relate to students' flexible use of writing styles? Is this driven by some executive component skill, or is this driven more broadly by students' prior knowledge and use of strategies? Studies aimed at answering these (and other) questions have the potential to provide crucial information about the role of flexibility in students' ability to produce high-quality text

CHAPTER 3

A MULTI-DIMENSIONAL ANALYSIS OF WRITING FLEXIBILITY IN AN AUTOMATED WRITING EVALUATION SYSTEM

In Chapter 2, we presented the linguistic flexibility hypothesis – the idea that skilled writing is related to a flexible use of linguistic style, rather than a static set of specific text properties (Allen, Snow, & McNamara, 2016). The results of this 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?

In the current study, we aim 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, and examine how this 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. Therefore, we examine whether students given access to spelling and grammar feedback during the writing process would produce texts that differed from their peers along the tested linguistic dimensions.

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 the results and our interpretations in light of prior research.

Automated Writing Evaluation

Researchers and educators have developed computer-based writing tools 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 (Allen, Jacovina, & McNamara, 2016). These tools have been developed with a variety of goals in mind (Dikli, 2006; Roscoe, Allen, Weston, Crossley, & McNamara, 2014; Weigle, 2013). For instance, automated essay scoring (AES) systems focus on the automatic scoring of students' essays and are typically employed by high-stakes testing companies to score the essay component of many standardized tests (Shermis & Burstein, 2003; 2013; Deane, 2013). These AES 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 (Dikli, 2006; Shermis & Burstein, 2003; 2013; Warschauer & Ware, 2006).

More recently, these AES systems have expanded beyond these assessment contexts and have been integrated with educational learning environments, such as automated writing evaluation (AWE) systems (Attali & Burstein, 2006; Crossley, Varner, Roscoe, & McNamara, 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 educational systems is to move these 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 face a wealth of criticism, which often centers 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 (Deane, 2013; Perelman, 2012). 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.

Current Study

In the current study, we examine essay writing in the context of an AWE system 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 prompts). Further, we examine whether these properties of student writing vary according to their literacy skills or with the presence of on-line low-level feedback. The students in this study wrote and revised six essays in the context of an AWE system that provided them 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. 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. Are the dimensions along which students vary the same when considering separate essay prompts as compared to drafts in response to the same prompts?
    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?

We first hypothesize that the developing writers in this study will exhibit stylistic flexibility (e.g., narrativity) across essay assignments, but predominantly 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.

Method

Participants

Participants (n = 131) in this study were high school students recruited from an urban environment located in the southwestern United States. All students were recruited from local high schools and publically posted flyers. These students were monetarily compensated for their participation in this experiment. On average, these participants 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 eleven 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 lasted between two and three weeks for each participant. During each session, participants wrote and revised two essays within the context of the AWE component of the Writing Pal (W-Pal), an intelligent tutoring system for writing instruction and practice (Roscoe & McNamara, 2013). In this AWE component of the system, participants had access to a word processor that prompted them to write an essay in response to an SAT-style prompt. All students were given 25 minutes to complete their initial essay draft, received automated 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.

Automated Essay Scoring and Feedback

During the study, students received both summative and formative feedback on their essays. The summative scores were driven by the W-Pal algorithm (McNamara, Crossley, Roscoe, Allen, & Dai, 2015), which calculates a variety of linguistic indices related to the submitted essay and provides both summative and formative feedback to student writers. The summative feedback delivered 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 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 revise their essays based on the feedback that they received.

Formative feedback is an important component of writing development, as it provides important knowledge to writers about components of high-quality writing. Additionally, formative feedback provides students with actionable recommendations for how to improve their writing, such as generating ideas and examples and maintaining cohesion through explicit text connections. The automated formative feedback in W-Pal was developed with this design in mind, and provides recommendations that relate to multiple writing strategies. Previous research evaluating the efficacy of the W-Pal system has found that this training results in improved essay scores, increased strategy knowledge, and improved revising strategies (Allen, Crossley, Snow, & McNamara, 2014; Allen, Crossley, Snow, Jacovina, Perret, & McNamara, 2015; Crossley, Varner, Roscoe, & McNamara, 2013).

Computational Text Analyses of Student Essays

To examine how students revised the texts they were assigned, the revised drafts were analyzed using Coh-Metrix. Coh-Metrix (McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014) is a computational text analysis tool that was developed to provide automated measures of text readability (Duran, Bellissens, Taylor, & McNamara, 2007). This tool analyzes texts at the word, sentence, and discourse levels to offer more nuanced information about the challenges and linguistic scaffolds contained within a given text. To account for the multiple dimensions of text readability, Graesser, McNamara, and Kulikowich (2011) developed the five Coh-Metrix Easability Components, which offer a detailed glance at the primary levels of text difficulty and are well aligned with an existing multilevel framework (Graesser & McNamara, 2011). These Easability Components relate to: *Narrativity, Word Concreteness, Syntactic Simplicity, Referential Cohesion,* and *Deep Cohesion*. In the current study, students' revised texts were analyzed along the five Easability Components produced by Coh-Metrix.

Reading Comprehension Assessment

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 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; Phillips, Norris, Osmond, & Maynard, 2002).

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 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.

Results

Percentage scores on the reading comprehension test suggest that participants 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 condition groups, we calculated a between-subjects ANOVA, which revealed that there were no significant differences between the reading scores for the participants 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.

Linguistic Flexibility Across Writing Assignments

We assessed the influence of prompt (essay writing assignment) and experimental condition (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 final 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; the details of the full model (Model 4) for each linguistic dimension are presented in tables in Appendix C. In these tables, 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.

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 shown in Table C.1. This suggests that, for some of the essay prompts, students' method of adapting their narrative style differed according to their reading comprehension skills.

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 (see Table C.2 for Model 4 details). 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 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 (see Table C.3).

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 overall 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 (see Table C.4) 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). 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 (see Table C.5 for Model 4 details).

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, higher skilled students tended to produce essays that were less narrative and referentially cohesive but more syntactically simple than their less skilled peers. Further, these reading 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 do suggest that the linguistic properties of student writing vary based on the prompt to which they are responding as well as individual differs 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 analyses 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 within their essays.

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 reading 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 final 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 of the models are presented below; the details of the full model (Model 3) for each dimension are presented in tables in Appendix D.

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 (Table C.6).

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 (Table C.7).

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 (Table C.8), suggesting that both more and less skilled students revised these words in similar ways.

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 (Table C.9).

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 (Table C.10).

Discussion. The results of our second set of analyses on students' 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.

Conclusion

In this study, we examined the relationship 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.

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 (Crossley, Kyle, Allen, & McNamara, 2014; Graham & Perin, 2007). 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.

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 in Chapter 3 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.

CHAPTER 4

AN EXAMINATION OF THE ONLINE BEHAVIORS UNDERLYING WRITING FLEXIBILITY

The studies presented in Chapters 2 and 3 of this dissertation take important steps toward developing a better understanding of linguistic flexibility, particularly as it manifests in the essays of developing writers. In these studies, we were primarily interested in understanding how these developing writers naturally varied the properties of their language in different contexts; therefore, we chose not to explicitly manipulate the audience or genre assigned to students in these studies. In Chapter 4, we build on these prior studies by conducting an analysis of linguistic flexibility that explicitly examines students' ability to respond to different audiences. In particular, this study prompted students to revise news articles so that they were more appropriate for different audiences.

The purpose of this final study is to understand whether students systematically adapt their language when they are explicitly instructed to write for audiences who can be assumed to have differing levels of prior knowledge and comprehension skills. Further, we aim to examine whether the linguistic changes that students make to the texts for these audiences reflect an accurate understanding of text readability across multiple dimensions.

Text Readability and Individual Differences

Students' prior knowledge is strongly related to their performance on academic writing tasks (Allen, Snow, Crossley, Jackson, & McNamara, 2014; McCutchen, 1986).

An individual’s *knowledge* can refer to an individual’s knowledge of writing itself (e.g., writing strategies, processes), as well as the domain knowledge required to complete a given assignment (e.g., science knowledge; Graham, Harris, & Mason, 2005). Further, students’ ability to *comprehend* texts is related to both their writing skills as well as their prior knowledge. This points to the existence of complex interactions among the knowledge and skills required to successfully produce texts for individual audiences.

A significant amount of research has been devoted to examining how text properties influence individuals’ processing and comprehension of texts. If such a strong relation between these text features and comprehension processes, how might individual differences in these cognitive skills and abilities relate to the ways in which individuals produce texts? Is it the case, for example, that students with lower reading comprehension skills produce texts that are easier to process and understand than students who have strong comprehension skills? Additionally, do student writers possess the knowledge to adapt their texts in ways that are appropriate for audiences who vary in their knowledge and skills?

Current Study

The purpose of this final dissertation study is to examine how student writers revise texts for audiences of different knowledge levels. In the previous studies, we examined how students varied the properties of their language in naturalistic educational writing contexts. Our interpretation of the results of these studies has been that this flexibility is related to an underlying understanding of the ways in which linguistic text features interact to influence readability overall. Thus, we have assumed that linguistic flexibility is an intentional and strategic behavior employed by skilled writers. However, we have not empirically tested this assumption.

The current study builds on this work by explicitly prompting students to revise texts for audiences of differing knowledge and literacy levels. Thus, the overall purpose of this study is to examine whether students adapt to these different audiences at all and, if so, whether they adapt in ways that are appropriate for the different audiences. Further, we examine whether this flexible adaptation to the different assigned audiences is related to the students' own comprehension skills.

Method

Participants

Participants (n = 95) in this study were undergraduate students from the Psychology 101 subject pool at Arizona State University. These participants were, on average, 19.2 years of age (range 17 to 24), 48% were female, 61% were Caucasian, 21% were Asian, 6% were Hispanic, and 12% reported other ethnicities. All students received course credit for completion of the study. Seven of the participants were dropped from the analyses because they misunderstood the instructions.

Study Procedure and Design

Participants in this study completed a comprehension assessment (described below) and then engaged in a set of text rating and revision tasks. They were given a general set of instructions for these tasks that explained that they would answer questions and revise texts to help ASU researchers understand how to develop texts that are appropriate for different audiences. They then completed the set of tasks, which consisted of 40 text ratings, wherein students rated the text based on the level they thought it is appropriate for (i.e., second grade through college), as well as their perceived understanding of the text (i.e., *I understand the text very well – I do not understand this text*).

For three texts distributed evenly throughout the 40 texts of this text set, students were asked to engage in an additional text *revision* task. After rating the texts along the same dimensions as the previous texts, they were presented with three blank text files and allotted 20 minutes to produce two new texts that render this original text appropriate for members of different audiences: a class of 4th grade students and a group of ASU professors. During this time, students' keystrokes and computer actions (e.g., copy, paste) were recorded. Following the text revision task, students will be asked to provide new ratings for each of the revised texts.

The three texts that students were asked to rate and revise throughout the study were simplified, non-academic news articles selected from the *Guardian Weekly*, a British-based publication with a wide international readership. In particular, the articles used in this experiment were taken from a corpus of *Guardian Weekly* news articles that have been revised such that they contain approximately 150 words and represent beginning and intermediate difficulty levels across 6 genres: business, culture, environment, politics, science, and world news (see Appendix B for example texts). This corpus of revised articles has been used in previous research to develop text readability measures (Allen, 2009; Crossley, Allen, & McNamara, 2012).

In this experiment, students were asked to rate texts from both the beginning and intermediate levels; however, the revision tasks only occurred for texts of intermediate grade level. To control for potentially complex interactions among text properties and individual differences, all students rated and revised texts in the same order.

Materials

Demographics questionnaire. Students' demographics were collected through a battery of self-report questions. These assessments relate to basic identifying information such as students' age, gender, and ethnicity.

Prior reading ability. The reading ability of the students 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 answering two to six questions about the content of the passage. These questions are 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; Phillips, Norris, Osmond, & Maynard, 2002).

Computational Analysis of Revised Texts

To examine how students revised the texts they were assigned, the revised drafts were analyzed using Coh-Metrix. Coh-Metrix (McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014) is a computational text analysis tool that was developed to provide automated measures of text readability (Duran, Bellissens, Taylor, & McNamara, 2007). This tool analyzes texts at the word, sentence, and discourse levels to offer more nuanced information about the challenges and linguistic scaffolds contained within a given text. To account for the multiple dimensions of text readability Graesser, McNamara, and Kulikowich (2011) developed the five Coh-Metrix Easability Components, which offer a detailed glance at the primary levels of text difficulty and are well aligned with an existing multilevel framework (Graesser & McNamara, 2011). These Easability Components relate to: Narrativity, Word Concreteness, Syntactic Simplicity, Referential Cohesion, and Deep Cohesion. In the current study, students' revised texts will be analyzed along the five Easability Components produced by Coh-Metrix.

Statistical Analyses

To examine whether students revised texts in ways that were meaningfully adapted to the two audiences, we conducted linear mixed-effects models using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015). In order to account for students' revisions to the texts, we first calculated difference scores between the Easability Component scores of the original texts and students' revised versions of these texts. Thus, a positive score indicated that the text was revised to be simpler, whereas a negative score indicates that the text was revised to be more difficult. These difference scores served as the dependent variables in our models. Because many of the students were unable to complete the third revision period due to wide variability in reading times, analyze were conducted solely on the revisions for the first two texts in this study.

As fixed effects in our models, we entered audience (professors were coded as -0.5 and students were coded as 0.5) and comprehension scores (grand mean centered) into the model, as well as an interaction term between these variables. As random effects, we included intercepts for the participants and the text they were asked to revise. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. For each of the models listed below, p-values were obtained by likelihood ratio tests of the full model compared to two reduced models. The null model included the random intercepts for each participant and text revision. Model 2 added the effect of audience and Model 3 added the effect of comprehension scores. The full model added the interaction term between comprehension scores and audience. Tables presented in Appendix D present the output of each model.

Results

Students' scores on the reading comprehension test suggested that they varied considerably in their literacy skills, ranging from a minimum score of 17.74% to a maximum score of 97.92% (*M* = 60.72, *SD* = 17.74).

Narrativity. Overall, students increased the narrativity of the texts by 6.58 points (*SD* = 14.73). In particular, the comparison between the null model and Model 2 indicated that there was a significant effect of audience on students' revisions of the texts' narrativity, χ^2 (1) = 68.176, $p < .001$, such that participants increased the narrativity of the texts for the class of fourth grade students (*M* = 12.48, *SD* = 16.98) more than for group of professors (*M* = 0.78, *SD* = 8.93). Adding the fixed effect of comprehension scores in Model 3 did not significant increase model fit ($p = 0.08$); however, the final model (Model 4) that included the interaction term did lead to a significant increase in model fit over Model 2, χ^2 (2) = 7.035, *p* < .05, indicating that students with higher reading comprehension scores were more likely to revise these texts in appropriate ways (i.e., revise them to be more narrative for the students) compared to the students who had lower reading comprehension scores (see Table D.1).

Syntactic Simplicity. Participants decreased the syntactic simplicity of the texts by 12.29 points (*SD* = 18.31). However, there were no significant effects of audience, suggesting that these revisions to the syntax did not differ for the fourth-grade students (*M* = -10.96, *SD* = 18.60) or professor (*M* = -13.60, *SD* = 17.97) texts. Further, there was no effect of reading ability on the syntactic changes that students made to the texts (*p* > .05). Thus, students did not significantly alter the syntactic complexity of the texts to be more appropriate for the group of fourth grade students or professors, nor did they alter the syntax differently based on their reading abilities (full model in Table D.2).

Word Concreteness. Participants, on average, increased the concreteness of the words in the texts during their revisions (*M* = 1.06, *SD* = 12.93). There was a significant main effect of audience for changes that students made to the word concreteness of the texts, χ^2 (1) = 13.058, *p* < .001. In particular, students tended to increase the concreteness of the words in the texts intended for the student audience (*M* = 3.60, *SD* = 13.94), but decrease the concreteness for the professor audience (*M* = -1.43, *SD* = 11.36). However, Models 3 and 4 did not significantly improve the fit of the model (*p* > .05), suggesting that these concreteness changes were not related to students' ability to comprehend texts (see Table D.3).

Referential Cohesion. Participants generally increased the referential cohesion when revising the texts (*M* = 18.35, *SD* = 24.20). There was a significant main effect of audience on the changes to the referential cohesion of the texts, χ^2 (1) = 46.851, $p < .001$, such that students increased the referential cohesion of the texts for the fourth-grade student audience (*M* = 25.53, *SD* = 25.99) significantly more than for the group of professors (*M* = 11.31, *SD* = 19.56). Adding the fixed effect of reading ability further improved model fit in Model 3, χ^2 (1) = 5.684, $p < .05$. The final model (Model 4) that included the interaction term led to a further increase in model fit, χ^2 (1) = 8.348, $p < .01$. This model (Table D.4) revealed that more skilled readers revised the referential cohesion of the texts differently than their less skilled peers, such that the student texts contained a higher degree of referential cohesion compared to the texts intended for the group of professors.

Deep Cohesion. On average, participants increased the deep cohesion of the texts during the revision period (*M* = 6.84, *SD* = 20.93). Further, there was a significant effect of audience on students' changes to the deep cohesion of the text, χ^2 (1) = 5.684, $p < .05$. The model indicated that students increased the deep cohesion of the texts more for the fourth-grade student audience (*M* = 10.50, *SD* = 26.18) compared to the professor audience (*M* = 3.24, *SD* = 13.10), thus appropriately increasing the readability of the text. Models 3 and 4 did not significantly improve the fit of the model, which suggests that the nature of these changes was not related to students' reading comprehension skills.

Conclusion

The final study in this dissertation examined whether students systematically revised texts when prompted to create new versions that were appropriate for audiences of different age levels and presumed reading skills. In particular, it examined whether these revisions led to increases in text readability for the audience of fourth-grade students as compared to the group of professors. Additionally, the study examined whether the nature of these text revisions interacted with students' own ability to comprehend texts.

Our predictions were largely confirmed by these analyses. In particular, the linguistic properties of students' text revisions systematically differed according to audience and were, for the most part, appropriate for the two audiences. The students tended to increase the readability of the new texts intended for the group of fourth-grade students, but not for the group of professors. These results suggest that student writers can engage in adaptive writing processes across multiple levels of the text and are at least somewhat aware of the scaffolds available in texts across these multiple levels.

The results of our analyses examining the effect of audience on text revisions reveal important information about students' understanding of text readability. Across four of the five dimensions (all dimensions except for syntactic simplicity), students revised texts so that they were easier to read for the group of elementary students compared to the group of professors. In particular, when students revised texts for the group of students, they used language that was more narrative and concrete, and they increased both the referential and deep cohesion. This finding is important for a number of reasons. First, it provides further confirmatory evidence for the assumption that providing an explicit audience in a writing task can dramatically alter the linguistic properties of the texts that students produce. Second, the results suggest that students are capable of revising texts in ways that are appropriate for different audiences. Thus, whether this knowledge is implicit or not, students seem to reflect an understanding of the role of linguistic features in the text comprehension process.

The current study additionally indicated that the nature of students' revisions to the text interacted with their reading abilities. For two of the linguistic dimensions (i.e., narrativity and referential cohesion), there were significant interactions between audience and reading ability. On both of these dimensions, students with higher comprehension scores generated revised versions of the texts that were more appropriate for the two audiences compared to the less skilled students. This suggests that students who had higher literacy skills were better able to engage in appropriate revisions at deeper levels of the text (i.e., narrativity and referential cohesion).

Overall, the current study takes an important step towards understanding the nature of students' linguistic flexibility by explicitly investigating how they revise previous written texts so that they are more appropriate for different audiences. These findings can strengthen our theoretical understanding of text production processes, as well as for discourse processes more broadly. By examining how individuals adapt their language for different groups of people, we can gain a better understanding of their linguistic knowledge and flexibility, as well as the role of perspective taking in the writing process. Additionally, results of this and future studies can be used to inform educational literacy interventions and tutoring systems. If we can identify when students are struggling to appropriately respond to different writing contexts (e.g., audiences, genres), educators may be able to use this information to provide more adaptive instruction and feedback to their students.

CHAPTER 5

GENERAL DISCUSSION

Successful writing results from a complex interaction of cognitive and social skills with the aim of generating texts that successfully convey meaning to others (Graham, 2006). This skill requires an individual to have developed multiple forms of knowledge (e.g., vocabulary, domain), cognitive skills (e.g., constructing sentences that follow grammatical rules), as well as the ability to strategically use language to connect and convey ideas in ways that are meaningful for particular audiences (Donovan & Smolkin, 2006; McNamara, 2013).

Importantly, communicating via text is a complex process to understand. Consequently, researchers and educators place a relatively weak emphasis on writing compared to other language and cognitive processes, such as reading or listening (Graham, Harris, & Santangelo, 2015; National Commission on Writing, 2004). Conducting research that examines the processes involved in *generating* texts presents significant and unique challenges, as there exists a wide amount of variability in the nature of writing tasks and the ways in which individuals can successfully communicate through text. For instance, imagine two individuals who are prompted to describe why it is important to maintain a positive attitude throughout life. One individual might rely on engaging narrative anecdotes, which draw in their readers and convince them to believe their argument. A different writer might rely on facts drawn from empirical research in the field of positive psychology. Although both of the produced texts successfully argue the same point, they have achieved this goal through widely different means.

The example above focuses on one particular area in which writing exhibits wide variability. However, there are many more such examples related to the contexts surrounding common writing tasks, such as differences in individuals' writing processes, expectations of specific genres, and researchers' varied metrics for "high-quality writing." In response to this complexity, researchers and educators have often developed assessments of writing proficiency that are highly de-contextualized and have little ecological communicative purpose. Although measures such as these can help to increase the reliability of writing research and standardized tests, they often end up reflecting constructs that are widely different from those that are experienced by individuals in real-world writing scenarios. It is rather difficult to imagine a scenario in which an individual would be asked to generate a text with no explicit purpose or audience. However, the majority of standardized tests and writing measures ask students to do just that – namely, students are expected to respond to prompts that are rarely given context or grounded in real-world problems.

These de-contextualized measurements can present serious problems for the valid study and assessment of the writing process. Recent research suggests that the properties of texts that students generate do not consistently relate to expert ratings of writing quality. For instance, the linguistic properties of high-quality writing have been shown to vary across different contexts, such as authors, grade levels, prompts, and contexts (Allen, Snow, & McNamara, 2016; Crossley, Roscoe, & McNamara, 2014; Varner, Roscoe, & McNamara, 2013). Therefore, it is possible that these product-based measures of writing quality may not adequately capture individuals’ writing proficiency, as they often miss out on the ways in which students are responding to demands of the task.

Researchers have recently hypothesized that a writers’ ability to flexibly adapt to these varied writing contexts may play an important role in their ability to produce high-quality texts (Allen, Snow, Crossley, Jackson, & McNamara, 2014). In particular, the degree to which an individual can change their language based on different contexts can potentially provide critical information about the writing process that moves beyond static measures of linguistic essay properties. The work presented in this dissertation builds on this proposition through analyses of naturalistic essay data across writing prompts (Chapter 2 and Chapter 3) and drafts (Chapter 3), as well as through analyses of students’ ability to revise texts appropriately for different audiences (Chapter 4).

The study presented in Chapter 2 proposed and tested the initial linguistic flexibility hypothesis – namely that writing skill is associated with an individual’s ability to flexibly employ linguistic properties rather than simply focus on their consistent use of a particular set of linguistic properties. This hypothesis was tested through analyses that leveraged both natural language processing and dynamic methodologies to model the variability in students’ use of narrative style across multiple essay prompts. The results from this study revealed that students who demonstrated greater flexibility in their use of narrativity across essays were also more likely to produce higher-quality essays. Additionally, the writers who demonstrated greater narrative flexibility also performed better on individual difference assessments related to their general literacy skills and prior world knowledge. These provided initial support for the linguistic flexibility hypothesis and revealed the potential benefits of analyzing the nature of linguistic variability to better understand the writing process.

Chapter 2 provided a strong foundation on which researchers could begin to examine the nature of flexibility within the context of the text production process. This study left a number of questions to be explored in future research, such as those related to the linguistic dimensions on which individuals demonstrate flexibility as well as whether individuals flexibly adapt in appropriate ways for different audiences. In Chapters 3 and 4, we built on this initial study by begin to address some of these unanswered questions. In particular, we analyzed students' writing at the word- (word concreteness), sentence- (syntactic simplicity), cohesion- (referential and deep cohesion), and stylistic- (narrativity) levels. The purpose of these multi-dimensional text analyses was to examine whether flexibility manifested in different ways across these text levels and whether these effects were related to students' literacy skills.

Chapter 3 examined this multi-dimensional linguistic flexibility in the context of an automated writing evaluation system. Across a series of models in this study, we found that the linguistic properties of students' essays significantly varied based across the individual prompts they responded to as well as across their original and revised drafts. In particular, there was a significant effect of prompt on all five linguistic dimensions and a significant effect of draft for all dimensions except for syntactic simplicity. These results indicate that students did not simply produce essays in the same way across essay prompts and drafts; instead, they flexibly adapted their language in response to the demands of the task. Further, the results of these multi-dimensional analyses suggest that students were sensitive to the properties of the texts across both surface- and deep-levels.

The results of Chapter 3 further indicated that providing students with spelling and grammar feedback had no impact on the properties of their writing, nor on their responses to the prompts or drafts. This finding makes sense in the context of prior research on lower-level mechanics feedback. Although research supports *teaching* mechanics to developing writers (Graham, 1983; Morris, Blanton, Blanton, & Perney, 1995), meta-analyses of writing instruction demonstrate that it is one of the least effective forms of writing interventions (Graham & Perin, 2007). Surveys of writing teachers suggest that a significant amount of classroom time is spent on grammar and spelling instruction (Cutler & Graham, 2008); however, the results from Chapter 3 in combination with prior research suggest that this time may be better spent on other aspects of the writing process.

In the final Chapter of this dissertation, we provided further support for these results. The purpose of this study was to examine whether students could appropriately revise texts for different audiences at multiple textual dimensions. Our results supported those from Chapter 3 and indicated that students' new versions of the texts were revised in ways that made them more appropriate for the fourth-grade students and for the group of professors. This is important because it indicates that students were, either implicitly or explicitly, aware of the linguistic scaffolds that are available to readers across different texts. They were able to use these linguistic properties to produce texts that were appropriately adapted for the individual audiences.

The results in the three chapters also provide important information about the relations between students' literacy skills and their flexible use of linguistic properties. The results of Chapter 2 revealed that students who had higher scores on a reading comprehension test also demonstrated more flexibility in their writing across prompts. However, in Chapter 3, our results indicated that there were no linguistic 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, and referential cohesion) along which reading ability interacted with some, but not all, of the prompts. Combined with the results from Chapter 2, 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). Chapter 4 provided further support for this interpretation, as students' revisions to the narrativity and referential cohesion of the texts demonstrated significant interactions with their reading abilities. These results suggest that students' own comprehension skills may help them to better understand how to scaffold the reading processes of others.

Limitations and Future Directions

Although the results presented in this dissertation are promising, there are a number of limitations to address in future research. First, the prompts that students were asked to respond to in Chapters 2 and 3 were relatively similar. Therefore, the type of flexibility that they 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 the study presented in Chapter 4 to address these limitations. In particular, studies will be conducted to examine how students adapt their language more explicitly when prompted to write for different audiences or for different purposes. In particular, we plan to 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 concern relates to our claims about the degree of flexibility that students demonstrate in our studies. 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 regarding this flexibility and this skill is no longer as important among more skilled writers. These and related questions remain to be answered in new 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 dissertation 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 discourse production 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.

REFERENCES

Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. System, 37, 585–99.

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

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: The Guilford Press.

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.

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.

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.

Applebee, A. N., Langer, J. A., Jenkins, L. B., Mullis, I., & Foertsch, M. A. (1990). *Learning to write in our nation's schools*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Arneodo, M., Arvidson, A., Badełek, B., Ballintijn, M., Baum, G., Beaufays, J., & Prytz, K. (1995). Measurement of the proton and the deuteron structure functions. Physics Letters B, 364, 107–115.

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

Barab, S. A., Gresalfi, M. S., Dodge, T., & Ingram-Goble, A. (2010). Narratizing disciplines and disciplinizing narratives: Games as 21st century curriculum. *International Journal for Gaming and Computer-Mediated Simulations, 2,* 17-30.

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

Benhamou, S., & Bovet, P, (1989). How animals use their environment: A new look at kinesis. *Animal Behavior, 38,* 375-383.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.

Bruner, J. (1986). Actual minds, possible worlds. Cambridge, MA: Harvard University Press.

Cheong, Y., & Young, R. M. (2006). A computational model of narrative generation for suspense. *Proceedings of AAAI 2006 Workshop on Computational Aesthetics.*

Collins, J. J., & De Luca, C. J. (1993). Open-loop and closed-loop control of posture: A random-walk analysis of centure-of-pressure trajectories. *Experimental Brain Research, 95,* 308–318.

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensive input: A case for intuitive approach. *Language Teaching and Research, 16,* 89-108.

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.

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T.F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. (pp. 1236-1231). Austin, TX: Cognitive Science Society.

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., Roscoe, R. D., McNamara, D. S., & Graesser, A. C. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 438-440). Auckland, New Zealand: AIED.

Crossley, S. A., Varner, 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 International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 202-207). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Using automated cohesion indices as a measure of writing growth in intelligent tutoring systems and automated essay writing systems. In K. Yacef et al. (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED)* (pp. 269-278). Heidelberg, Berlin: Springer.

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.

Crowhurst, M. (1990). Reading/writing relationships: An intervention study. *Canadian Journal of Education, 15*, 155-172.

Cutler, L., & Graham, S. (2008). Primary grade writing instruction: A national survey. *Journal of Educational Psychology, 100,* 907-919.

Davison, A. (1984). Readability formulas and comprehension. In G. G. Duffy, L. R. Roehler, R. Mason, & J. Mason (Eds.), *Comprehension instruction: Perspectives and suggestions* (pp. 128-143). New York: Longman.

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.

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.

Duran, N., Bellissens, C., Taylor, R., & McNamara, D. (2007). Qualifying text difficulty with automated indices of cohesion and semantics. In D.S. McNamara and G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 233-238). Austin, TX: Cognitive Science Society.

Ferrari, M., Bouffard, T., & Rainville, L. (1998). What makes a good writer? Differences in good and poor writers' self-regulation of writing. *Instructional Science, 26*, 473-488. doi:10.1023/A:1003202412203

Fitzgerald, J. & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist*, *35,* 39-50.

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

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.

Gernsbacher, M. A. (1997). Coherence cues mapping during comprehension. In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships in the production and comprehension of text* (pp. 3-21). Mahwah, NJ: Erlbaum.

Graesser, A. C., Cai, Z., Louwerse, M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that helps survey methodologists improve the comprehensibility of questions. *Public Opinion Quarterly*, *70*, 3–22.

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.

Graesser, A. C., Olde, B., & Klettke, B. (2002). How does the mind construct and represent stories? In M. C. Green, J. J. Strange, & T. C. Brock (Eds.), *Narrative impact: Social and cognitive foundations* (pp. 231-263). Mahwah, NJ: Lawrence Erlbaum Associates.

Graham, S. (1983). Effective spelling instruction. *Elementary School Journal, 83,* 560-567.

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.

Graham, S., Bollinger, A., Olson, C. B., D'Aoust, C., MacArthur, C., McCutchen, D., & Ollinghouse, N. (2012). *Teaching elementary school students to be effective writers – An educator's practice guide for the Institute of Education Sciences.* United States Department of Education.

Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30*, 207-241.

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

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

Haberlandt, K., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114,* 357–374.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

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

Hiebert, E. H. (2002). Standards, assessments and text difficulty. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 337–369). Newark, DE: International Reading Association.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

Huot, B. (1996). Towards a new theory of writing assessment. *College Composition and Communication, 47*, 549–566.

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

Kincaid, J., Fishburne, R., Rogers, R., & Chissom, B. (1975). *Derivation of new readability formulas for navy enlisted personnel.* Branch Report 8–75. Millington, TN: Chief of Naval Training.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review, 95,* 163–182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, England: Cambridge University Press.

Kintsch, W. & van Djik, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

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

Lobry, J. R., (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biological Evolution, 13,* 660–665.

Longo, B. (1994). The role of metadiscourse in persuasion. *Technical Communication, 41,* 348–352.

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

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25,* 431-444.

McCutchen, D. (2000). Knowledge, processing, and working memory: Implication for a theory of writing. *Educational Psychologist, 35,* 13-23.

McCutchen, D., Covill, A., Hoyne, S. H., & Mildes, K. (1994). Individual differences in writing: Implications of translating fluency. *Journal of Educational Psychology, 86,* 256–266.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1-30.

McNamara, D. S. (2013). The epistemic stance between the author and the reader: A driving force in the cohesion of text and writing. *Discourse Studies, 15,* 575-592.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27,* 57-86.

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

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.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188-205). Hershey, PA: IGI Global.

McNamara, D.S., Graesser, A.C., & Louwerse, M.M. (2012). Sources of text difficulty: Across genres and grades. In J.P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89-116). Lanham, MD: R&L Education

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

McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43.

McNamara, D. S. & Magliano, J. P. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation.* New York, NY: Elsevier Science.

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34*, 147-171.

Meadows, M., & Billington, L. (2005). *Review of the literature on marking reliability*. Report for the Qualifications and Curriculum Authority. London: National Assessment Agency.

Millis, K. K., Magliano, J. P., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading, 10,* 251–283.

Morris, D., Blanton, L., Blanton, W., & Perney, J. (1995). Spelling instruction and achievement in six classrooms. *Elementary School Journal, 96*, 145–162.

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

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

Nelson, C. R., & Plosser, C. R., (1982). Trends and random walks in macroeconmic time series: some evidence and implications. *Journal of Monetary Economics, 10,* 139-162.

Newkirk, T. (1997). *The performance of self in student writing.* Portsmouth, NH: Heinemann-Boynton/Cook.

Oakhill, J., & Yuill, N. (1996). Higher order factors in comprehension disability: Processes and remediation. In C. Cornaldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 69–72). Mahwah, NJ: Erlbaum.

Olinghouse, N. G., Grahan, S., & Gillespie, A. (2015). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology, 107,* 391-406.

O'Reilly, T., Best, R., & McNamara, D. S. (2004). Self-explanation reading training: Effects for low-knowledge readers. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Cognitive Science Society* (pp. 1053-1058). Mahwah, NJ: Erlbaum.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43,* 121-152.

O'Reilly, T., Taylor, R. S., & McNamara, D. S. (2006). Classroom based reading strategy training: Self-explanation vs. reading control. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1887). Mahwah, NJ: Erlbaum.

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

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.

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.

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

Rafoth, B., & Rubin, D. L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication, 1,* 446-458.

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., Crossley, S. A., Snow, E. L., Varner, L. K., & McNamara, D. S. (2014). Writing quality, knowledge, and comprehension correlates of human and automated essay scoring. In W. Eberle & C. Boonthum-Denecke (Eds.), *Proceedings of the 27th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 393-398). Palo Alto, CA: AAAI Press.

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.

Saddler, B. & Graham, S. (2007). The relationship between writing knowledge and writing performance among more and less skilled writers. *Reading and Writing Quarterly, 23,* 3, 231-248.

Schank, R. C., & Abelson, R. P. (1995). Knowledge and memory: The real story. In R. S. Wyer (Ed.), *Knowledge and memory: The real story* (pp. 1-85). Hillsdale, NJ: Lawrence Erlbaum Associates.

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) NY; Guilford.

Shanahan, T., & Tierney, R. J. (1990). Reading-writing relationships: Three perspectives. In J. Zutell & S. McCormick (Eds.), *Literacy theory and research: Analyses from multiple paradigms* (Thirty-ninth yearbook of the National Reading Conference, pp. 13-34). Chicago: National Reading Conference.

Shermis, M. D., & 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.

Snow, E. L., Allen L. K., Russell, D. G., & McNamara, D. S. (2014). Who's in control?: Categorizing nuanced patterns of behaviors within a game-based intelligent tutoring system. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 185-192). London, England.

Snow, E. L., Likens, A., Jackson, G. T., & McNamara, D. S. (2013). Students' walk through tutoring: Using a random walk analysis to profile students. *Proceedings of the 2013 Educational Data Mining Conference*. Berlin / Heidelberg, Germany: Springer.

Soller, A., & Lesgold, A. (2003). A computational approach to analyzing online knowledge sharing interaction. In *Proceedings of the 11th Annual Meeting of Artificial Intelligence in Education* (pp. 253-260). Sydney, Australia: Springer.

Swanson, H. L., & Berninger, V. W. (1996). Individual differences in children's working memory and writing skill. *Journal of Experimental Child Psychology, 63*, 358-385.

Tierney, R. J., & Shanahan, T. (1991). Research on the reading-writing relationship: Interactions, transactions, and outcomes. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *The handbook of reading research* (Vol. 2, pp. 246-280). New York: Longman.

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.

Vorderer, P., Wulff, H. J., & Friedrichsen, M. (1996). *Suspense: Conceptualizations, theoretical analyses, and empirical explorations.* Mahwah, NJ: Erlbaum.

Warschauer, M., and 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*, 36-54.

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

Zhou, M. (2013). Using traces to investigate self-regulatory activities: A study of self-regulation and achievement goal profiles in the context of web search for academic tasks. *Journal of Cognitive Education and Psychology, 12,* 287-305.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123,* 162–185.

APPENDIX A

PRETEST AND POSTTEST ESSAY PROMPTS

Essay Prompt 1. You will now have 25 minutes 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.

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?

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.

Essay Prompt 2. You will now have 25 minutes 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.

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?

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.

APPENDIX B

EXAMPLE TEXTS TO RATE AND REVISE

Beginning Level Text. Last month the senior elephant keeper at London Zoo, Jim Robson, was killed by one of the elephants he loved. Robson had worked at the zoo for 26 years, the past 16 in the elephant house. He was crushed to death by the elephant in front of about 100 people. It was not funny.

This was a tragic death, and it could be the beginning of the end of London Zoo - perhaps of all Britain's urban zoos. Last week the zoo announced that its three elephants were to be moved to Whipsnade wild animal park, a country park outside London. The zoo's director-general, Michael Dixon, in the statement. "We will be sorry to see the elephants go; there have been elephants in London Zoo since 1831."

One newspaper article said that this was a crisis for the zoo, and for all zoos, because if London Zoo admits that it cannot keep "charismatic megaspecies", it is accepting that it has no future. Many smaller zoo animals are wonderful, but they will not attract large numbers of visitors to the zoo. Lions, tigers, gorillas, giraffes, pandas, rhinos - and most of all elephants - are what makes a visit to the zoo memorable.

Intermediate Level Text. It may not make all parents jump for joy but a report published today shows the favourite reading material of young teenagers is Heat magazine. Parents may be more pleased to see that Anne Frank's diary, books by Anthony Horowitz and CS Lewis' The Lion, the Witch and the Wardrobe are also in the top ten.

The celebrity gossip and news magazine comes top when 11 to 14-year-olds are asked to name their favorite read, followed by teenage girls' magazine Bliss, which comes joint second with reading song lyrics online. They are followed by reading computer game cheats advice online, and then reading your own blog or fan fiction.

The first books in the list are the Harry Potter series at number five. Proving how inconsistent teenagers are, Harry Potter is also number eight in the most hated reading material top ten.

The results are in a report called Read Up, Fed Up: Exploring Teenage Reading Habits in the UK Today, which was commissioned by organizers of the National Year of Reading, which Gordon Brown launched in January.

Other books on the favorites list are Anne Frank's diary at number six, Anthony Horowitz novels at eight, the CS Lewis classic at number nine and books by Louise Rennison, author of the Confessions of Georgia Nicolson series, in joint tenth place with BBC Online.

APPENDIX C

LINEAR MIXED-EFFECTS MODELS FROM CHAPTER 3

Table C.1.

Full model including prompt, reading ability, and prompt by reading ability interaction to predict narrativity

	Narrativity		
	B	*CI*	*p*
Fixed Parts			
(Intercept): Competition Prompt	71.80	68.59 – 75.01	<.001
Prompt: Loyalty	6.22	2.77 – 9.68	**<.001**
Prompt: Images	-2.09	-5.53 – 1.35	.235
Prompt: Memories	17.72	14.26 – 21.18	**<.001**
Prompt: Patience	8.08	4.62 – 11.54	**<.001**
Prompt: Winning	7.41	3.95 – 10.87	**<.001**
Reading Ability	-0.41	-0.57 – -0.25	**<.001**
Prompt: Loyalty * Reading Ability	0.22	0.04 – 0.39	**.016**
Prompt: Images * Reading Ability	0.13	-0.05 – 0.30	.148
Prompt: Memories * Reading Ability	0.25	0.08 – 0.43	**.005**
Prompt: Patience * Reading Ability	0.11	-0.07 – 0.28	.230
Prompt: Winning * Reading Ability	0.10	-0.07 – 0.27	.263
Random Parts			
σ^2		183.151	
$\tau_{00, ID}$		135.656	
N_{ID}		119	
ICC_{ID}		0.426	
Observations		706	
R^2 / Ω_0^2		.605 / .596	

Table C.2.

Full model including prompt, reading ability, and prompt by reading ability interaction to predict syntactic simplicity

	Syntactic Simplicity		
	B	*CI*	*p*
Fixed Parts			
(Intercept): Competition Prompt	47.43	43.30 – 51.55	<.001
Prompt: Loyalty	-4.21	-8.45 – 0.04	.053
Prompt: Images	-8.33	-12.56 – -4.11	**<.001**
Prompt: Memories	-12.56	-16.80 – -8.31	**<.001**
Prompt: Patience	-5.97	-10.22 – -1.72	**.006**
Prompt: Winning	4.35	0.10 – 8.60	**.045**
Reading Ability	0.25	0.04 – 0.46	**.019**
Prompt: Loyalty * Reading Ability	-0.27	-0.49 – -0.06	**.013**
Prompt: Images * Reading Ability	-0.03	-0.25 – 0.18	.750
Prompt: Memories * Reading Ability	-0.18	-0.39 – 0.03	.101
Prompt: Patience * Reading Ability	-0.02	-0.23 – 0.20	.864
Prompt: Winning * Reading Ability	-0.04	-0.26 – 0.17	.696
Random Parts			
σ^2		276.412	
$\tau_{00, ID}$		250.143	
N_{ID}		119	
ICC_{ID}		0.475	
Observations		706	
R^2 / Ω_0^2		.596 / .585	

Table C.3.

Full model including prompt, reading ability, and prompt by reading ability interaction to predict word concreteness

	Word Concreteness		
	B	*CI*	*p*
Fixed Parts			
(Intercept): Competition Prompt	23.67	19.95 – 27.38	<.001
Prompt: Loyalty	4.67	-0.23 – 9.56	.062
Prompt: Images	-8.56	-13.46 – -3.66	**<.001**
Prompt: Memories	15.93	11.06 – 20.81	**<.001**
Prompt: Patience	-5.38	-10.28 – -0.48	**.032**
Prompt: Winning	-0.31	-5.21 – 4.58	.900
Reading Ability	0.14	-0.04 – 0.33	.131
Prompt: Loyalty * Reading Ability	-0.13	-0.37 – 0.12	.318
Prompt: Images * Reading Ability	-0.01	-0.26 – 0.24	.928
Prompt: Memories * Reading Ability	-0.10	-0.35 – 0.15	.425
Prompt: Patience * Reading Ability	0.00	-0.25 – 0.25	.992
Prompt: Winning * Reading Ability	-0.09	-0.33 – 0.16	.500
Random Parts			
σ^2		368.082	
$\tau_{00, ID}$		58.580	
N_{ID}		119	
ICC_{ID}		0.137	
Observations		706	
R^2 / Ω_0^2		.333 / .314	

Table C.4.

Full model including prompt, reading ability, and prompt by reading ability interaction to predict referential cohesion

	Referential Cohesion		
	B	*CI*	*p*
Fixed Parts			
(Intercept): Competition Prompt	59.90	55.17 – 64.62	<.001
Prompt: Loyalty	15.50	10.65 – 20.34	**<.001**
Prompt: Images	8.45	3.60 – 13.29	**<.001**
Prompt: Memories	-8.63	-13.45 – -3.81	**<.001**
Prompt: Patience	-2.02	-6.87 – 2.82	.414
Prompt: Winning	-4.34	-9.19 – 0.51	.080
Reading Ability	-0.50	-0.74 – -0.26	**<.001**
Prompt: Loyalty * Reading Ability	0.13	-0.12 – 0.37	.303
Prompt: Images * Reading Ability	0.27	0.03 – 0.51	**.031**
Prompt: Memories * Reading Ability	0.17	-0.07 – 0.42	.161
Prompt: Patience * Reading Ability	0.00	-0.24 – 0.25	.992
Prompt: Winning * Reading Ability	0.12	-0.13 – 0.36	.346
Random Parts			
σ^2		359.804	
$\tau_{00, ID}$		332.230	
N_{ID}		119	
ICC_{ID}		0.480	
Observations		706	
R^2 / Ω_0^2		.631 / .623	

Table C.5.

Full model including prompt, reading ability, and prompt by reading ability interaction to predict deep cohesion

	Deep Cohesion		
	B	*CI*	*p*
Fixed Parts			
(Intercept): Competition Prompt	84.88	81.34 – 88.43	<.001
Prompt: Loyalty	-9.69	-14.47 – -4.91	**<.001**
Prompt: Images	-3.84	-8.61 – 0.94	.116
Prompt: Memories	-1.52	-6.28 – 3.24	.531
Prompt: Patience	7.12	2.34 – 11.90	**.004**
Prompt: Winning	0.07	-4.71 – 4.85	.978
Reading Ability	0.06	-0.12 – 0.24	.492
Prompt: Loyalty * Reading Ability	-0.05	-0.29 – 0.19	.669
Prompt: Images * Reading Ability	-0.18	-0.43 – 0.06	.133
Prompt: Memories * Reading Ability	-0.11	-0.35 – 0.13	.383
Prompt: Patience * Reading Ability	-0.07	-0.31 – 0.17	.569
Prompt: Winning * Reading Ability	-0.13	-0.37 – 0.11	.289
Random Parts			
σ^2		350.525	
$\tau_{00, ID}$		38.955	
N_{ID}		119	
ICC_{ID}		0.100	
Observations		706	
R^2 / Ω_0^2		.245 / .214	

Table C.6.

Full model including reading ability, draft, and draft by reading ability interaction to predict narrativity

	Narrativity		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	82.76	80.60 – 84.92	<.001
Reading Ability	-0.22	-0.33 – -0.12	**<.001**
Draft	0.48	0.03 – 0.93	**.037**
Draft * Reading Ability	-0.01	-0.03 – 0.02	.577
Random Parts			
σ^2		18.793	
$\tau_{00, ID}$		464.906	
ρ_{01}		-0.316	
N_{ID}		119	
ICC_{ID}		0.961	
Observations		1411	
R^2 / Ω_0^2		.975 / .975	

Table C.7.

Full model including reading ability, draft, and draft by reading ability interaction to predict syntactic simplicity

	Syntactic Simplicity		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	39.94	36.90 – 42.98	<.001
Reading Ability	0.13	-0.02 – 0.29	.089
Draft	0.36	-0.24 – 0.97	.235
Draft * Reading Ability	-0.00	-0.03 – 0.03	.778
Random Parts			
σ^2		33.103	
$\tau_{00, ID}$		579.485	
ρ_{01}		-0.604	
N_{ID}		119	
ICC_{ID}		0.946	
Observations		1411	
R^2 / Ω_0^2		.970 / .970	

Table C.8.

Full model including reading ability, draft, and draft by reading ability interaction to predict word concreteness

	Word Concreteness		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	22.54	20.63 – 24.46	<.001
Reading Ability	0.10	0.00 – 0.19	**.050**
Draft	-0.74	-1.38 – -0.11	**.023**
Draft * Reading Ability	0.03	-0.01 – 0.06	.127
Random Parts			
σ^2		37.185	
$\tau_{00, ID}$		350.815	
ρ_{01}		-0.377	
N_{ID}		119	
ICC_{ID}		0.904	
Observations		1411	
R^2 / Ω_0^2		.960 / .958	

Table C.9.
Full model including reading ability, draft, and draft by reading ability interaction to predict referential cohesion

	Referential Cohesion		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	65.60	62.17 – 69.03	<.001
Reading Ability	-0.38	-0.56 – -0.21	**<.001**
Draft	1.00	0.34 – 1.66	**.003**
Draft * Reading Ability	-0.00	-0.04 – 0.03	.816
Random Parts			
σ^2		39.684	
$\tau_{00, ID}$		708.659	
ρ_{01}		-0.253	
N_{ID}		119	
ICC_{ID}		0.947	
Observations		1411	
R^2 / Ω_0^2		.974 / .974	

Table C.10.

Full model including reading ability, draft, and draft by reading ability interaction to predict deep cohesion

	Deep Cohesion		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	86.92	85.36 – 88.48	<.001
Reading Ability	-0.03	-0.11 – 0.05	.476
Draft	0.71	0.09 – 1.33	**.024**
Draft * Reading Ability	0.02	-0.01 – 0.05	.164
Random Parts			
σ^2		35.345	
$\tau_{00, ID}$		391.664	
ρ_{01}		-0.660	
N_{ID}		119	
ICC_{ID}		0.917	
Observations		1411	
R^2 / Ω_0^2		.956 / .954	

APPENDIX D

LINEAR MIXED-EFFECTS MODELS FROM CHAPTER 4

Table D.1.
Full model including audience, reading ability, and audience by reading ability interaction to predict narrativity change

	Narrativity Difference		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	0.74	-6.87 – 8.34	.865
Audience	11.63	9.07 – 14.18	**<.001**
Reading Ability	-0.00	-0.11 – 0.11	.989
Audience * Reading Ability	0.15	0.00 – 0.29	**.047**
Random Parts			
σ^2		142.948	
$\tau_{00, ID}$		7.382	
$\tau_{00, TextRevisionName}$		28.255	
N_{ID}		87	
$N_{TextRevisionName}$		2	
ICC_{ID}		0.041	
$ICC_{TextRevisionName}$		0.158	
Observations		337	
R^2 / Ω_0^2		.374 / .371	

Table D.2.

Full model including audience, reading ability, and audience by reading ability interaction to predict syntactic simplicity change

	Syntactic Simplicity Change		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	-13.68	-16.77 – -10.60	<.001
Audience	2.79	-0.51 – 6.08	.098
Reading Ability	-0.10	-0.28 – 0.07	.256
Audience * Reading Ability	0.05	-0.14 – 0.23	.617
Random Parts			
σ^2		236.775	
$\tau_{00, \text{ID}}$		93.882	
$\tau_{00, \text{TextRevisionName}}$		0.000	
N_{ID}		87	
$N_{\text{TextRevisionName}}$		2	
ICC_{ID}		0.284	
$ICC_{\text{TextRevisionName}}$		0.000	
Observations		337	
R^2 / $\Omega_0{}^2$		.469 / .402	

Table D.3.
Full model including audience, reading ability, and audience by reading ability interaction to predict word concreteness change

	Word Concreteness Change		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	-1.43	-3.34 – 0.47	.141
Audience	5.04	2.34 – 7.75	**<.001**
Reading Ability	-0.01	-0.12 – 0.10	.827
Audience * Reading Ability	0.04	-0.12 – 0.19	.627
Random Parts			
σ^2		160.345	
$\tau_{00, \text{ID}}$		0.000	
$\tau_{00, \text{TextRevisionName}}$		0.000	
N_{ID}		87	
$N_{\text{TextRevisionName}}$		2	
ICC_{ID}		0.000	
$\text{ICC}_{\text{TextRevisionName}}$		0.000	
Observations		337	
R^2 / Ω_0^2		.039 / .039	

Table D.4.
Full model including audience, reading ability, and audience by reading ability interaction to predict referential cohesion change

	Referential Cohesion Change		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	6.72	-8.35 – 21.80	.388
Audience	-5.22	-18.71 – 8.26	.449
Reading Ability	7.81	-13.65 – 29.27	.477
Audience * Reading Ability	31.78	10.41 – 53.15	**.004**
Random Parts			

σ^2	309.699
$\tau_{00, \text{ID}}$	168.380
$\tau_{00, \text{TextRevisionName}}$	22.661
N_{ID}	87
$N_{\text{TextRevisionName}}$	2
ICC_{ID}	0.336
$\text{ICC}_{\text{TextRevisionName}}$	0.045
Observations	337
R^2 / Ω_0^2	.582 / .557

Table D.5.

Best fitting linear mixed-effect model predicting deep cohesion change

	Deep Cohesion Change		
	B	*CI*	*p*
Fixed Parts			
(Intercept)	3.21	-4.26 – 10.68	.476
Audience	7.12	2.91 – 11.34	**.001**
Reading Ability	-0.00	-0.18 – 0.17	.956
Audience * Reading Ability	-0.11	-0.35 – 0.13	.384
Random Parts			
σ^2		389.367	
$\tau_{00, \text{ID}}$		8.448	
$\tau_{00, \text{TextRevisionName}}$		24.282	
N_{ID}		87	
$N_{\text{TextRevisionName}}$		2	
ICC_{ID}		0.020	
$ICC_{\text{TextRevisionName}}$		0.058	
Observations		337	
R^2 / Ω_0^2		.137 / .131	