Research Article

Across Performance Contexts: Using Automated Writing Evaluation to Explore Student Writing

Daniel F. McCaffrey, *Educational Testing Service*

Mo Zhang, *Educational Testing Service*

Jill Burstein, *Duolingo, Inc*[1]

Structured Abstract

- **Background:** This exploratory writing analytics study uses argumentative writing samples from two performance contexts—standardized writing assessments and university English course writing assignments—to compare (1) linguistic features in argumentative writing and (2) relationships between linguistic characteristics and academic performance outcomes. Writing data from this study come from 180 students enrolled at five four-year universities in the United States. Automated writing evaluation (AWE) tools were used to generate linguistic features from students' writing.
- **Literature Review:** Few studies have been conducted that use AWE to examine postsecondary writing skill and relationships between writing and broader academic performance outcomes. To address this gap, our study draws upon research on standardized and coursework writing, construct modeling, and AWE feature design. We also draw on related work to demonstrate how AWE can provide insights about linguistic characteristics of students' writing and relationships of that writing to academic performance factors.

[1] Jill Burstein completed her work on this paper while employed at ETS.

- **Research Questions:** We examine the feasibility of using AWE to examine writing samples from standardized assessment and coursework contexts to assess how the samples compare to each other and how they are related to measures of students' academic performance. We organize this study around two research questions (RQ):

  RQ1: What can AWE tell us about variation in student writing samples drawn from two specific performance contexts: standardized argumentative writing versus coursework argumentative writing?

  RQ2: What are the relationships between writing subconstructs (as captured by AWE) and other measures of students' academic performance?

- **Methodology:** As part of a larger program of research (n = 735), study participants (n = 180) were administered two standardized test instruments: the HEIghten Written Communication (WC) test and the HEIghten Critical Thinking test. The HEIghten tests are commercially available tests designed for assessing postsecondary students' skills. Students also submitted writing samples from their coursework writing assignments. We collected institutional data, including SAT total scores, SAT Reading scores, ACT total scores, high school GPAs, and college GPAs. AWE tools were used to generate the 36 writing features—aggregated to create six subconstruct composite scores—that were used to analyze both the HEIghten WC test and the coursework writing samples. To answer RQ1, we used multiple statistical analysis and data visualization methods. The distributions of the AWE subconstruct scores between the HEIghten WC and coursework responses were compared in histograms and boxplots. Paired t tests were conducted for each composite feature to examine whether the means of the composite scores were statistically the same (H0). To answer RQ2, we used the academic outcome variables from the institutional data noted previously. We also examined the correlations of AWE feature scores with the scores students received from the multiple-choice portion of the HEIghten WC test.

- **Results:** Findings showed that AWE feature distributions differed between the two performance contexts. This finding suggests that both AWE feature interpretation and the relationship of these features to academic performance factors are associated with the writing context. Findings also suggest that writing data from both contexts offers a complementary picture of students' writing achievement and relationships with performance outcomes.

- **Discussion:** Using AWE feature interpretation, this study establishes a foundation for variation in postsecondary student writing and demonstrates how a granular

sense of variation can be leveraged in order to understand relationships between writing achievement and broader postsecondary academic performance factors.

- **Conclusions:** Study findings have practical implications for how institutions might use writing analytics to obtain detailed information about student writing and, in turn, relate that information to other performance measures to provide relevant support for students. However, study findings must be qualified in terms of study sample, genres examined, standardized criterion measures, and shifting educational pedagogies.

Keywords: writing analytics, postsecondary education, standardized writing assessment, coursework writing

1.0 Background

1.1 Study Motivation

In the United States, writing is viewed as key to educational and workplace success (Allan & Driscoll, 2014; Arum & Roska, 2011; Kelly-Riley, 2015). Despite how important writing skills are, there is concern that many students do not develop the writing domain knowledge required for school, college, and workplace success. In terms of kindergarten through high school learning, Graham and Harris (2014) point out that many students do not develop writing skills in K-12 classrooms at levels which suggest competency. The National Assessment of Educational Progress (NAEP) has demonstrated that 54 percent of 8th graders and 52 percent of 12th graders performed at the Basic level (denoting only partial mastery of the prerequisite knowledge and skills that are fundamental for proficient work) in writing in 2011. Conversely, only three percent of 8th and 12th graders in 2011 performed at the Advanced level (denoting superior performance) in writing (National Center for Education Statistics, 2012). This performance defecit is compounded by studies of K-12 writing instruction that reveal that little time is devoted to teaching writing (Applebee & Langer, 2011). In postsecondary writing, similar deficits have also been identified. In an analysis of the VALUE rubric collaborative sponsored by the Association of American Colleges and Universities, Rhodes (2021) reports results from a study involving over 29,000 samples of college student writing. In this analysis, student writing was evaluated as to whether it demonstrated the level of achievement expected for students who had completed at least 75 percent of the requirements for the baccalaureate degree. The percentage of students meeting this level of achievement (called the capstone level to indicate performance expected for students approaching degree attainment) for defined aspects of writing are as follows: context/purpose = 18 percent, content development = 13 percent, genre/conventions = 10 percent, sources/evidence = 12 percent, and style/mechanics = 10 percent. Far lower percentages of students from two-year institutions met these levels: context/purpose = eight percent, content

development = five percent, genre/conventions = four percent, sources/evidence = four percent, and style/mechanics = two percent. Employers have also observed that many workers do not possess relevant writing skills (Hart Research Associates, 2015), and this skills deficit may lead to revenue loss due to communication failure (Bolchover, 2012).

In the context of these challenges, our research investigates the relationships between linguistic features in a given discourse mode (argumentation) across two settings typically encountered by U.S. postsecondary writing students (standardized tests and classroom writing). We then examine relationships between these features and measures of academic performance. As a descriptive study, our study motivation is to continue to gain insights about the nature of writing proficiency—seen through the lens of linguistic features—in academic performance.

1.2 Writing Contexts: Standardized Writing and Coursework Writing

Both standardized and coursework writing may usefully be understood as responses to constructed response tasks (Baldwin et al., 2005). As opposed to multiple choice tasks, constructed response tasks are designed to invite students to create—that is, construct—a response to a given writing task (Bennett & Ward, 2009). In some cases, those tasks may be very basic, requiring a very limited genre (such as a brief essay), a basic mode of discourse (narration), and a foundational set of writing skills (with emphasis on example development). In other cases, those tasks may be expansive, requiring complex genres (source-based writing), advanced discourse modes (argumentation), and a complex set of writing knowledge, skills, and attitudes (with emphasis on genre-related language use and conscious reflection). When a student writes to a standardized task, that task is designed to maintain a consistent environment adhering to predetermined rules and specifications, and results are reported according to a defined scale using norms based on a representative sample of students (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). When a student writes to a classroom task, those standards are not necessarily in force. In his study of 2,202 writing assignments taken from 400 courses across the disciplines in a wide variety of 100 U.S. postsecondary institutions, Melzer (2014) found that students wrote to tasks of limited purpose in which they reported factual details from readings and lectures. In terms of length, the short-answer examination was the dominant genre; in terms of assessment, the majority of instructors focused on grammatical correctness even when the syllabus claimed their courses focused on critical thinking.

Currently, little is known about how or if inferences about student ability derived from standardized writing align with inferences drawn from writing in other contexts such as coursework. We have very little understanding about how writers perform on specific features of the writing construct (such as those shown in Table 1) when writing in different performance contexts. In general, a more comprehensive review of the ways that information about writing differs between standardized assessment and classroom coursework writing could inform the role that each mode of evaluation can play in our understanding of students' writing competency.

1.3 Writing Instruction: Modeling Writing Constructs

Influencing both standardized and classroom contexts, evidence-based models exist for writing instruction that result in the ability of students to perform competently in both standardized and classroom constructed response tasks. In U.S. postsecondary writing, special attention is paid to the interface between secondary and college writing due, in a large measure, to the near universal existence of first-year general education writing requirements for degree attainment. The federally sponsored Institute of Education Sciences (IES) publishes practice guides to provide educators with panel-based guidelines to address specific areas of instruction. These practice guidelines—developed and published in conjunction with experts according to a defined level of evidence model—resulted in *Teaching Secondary Students to Write Effectively* (Graham et al., 2016). The IES guidelines recommend the following three pedagogies: explicit instruction in appropriate writing strategies using a model-practice-reflect instructional cycle, integration of writing and reading to emphasize key writing features, and use of assessments of student writing to inform instruction and feedback. On the postsecondary level, professional organizations of national influence—the Council of Writing Program Administrators (CWPA), National Council of Teachers of English (NCTE), and National Writing Project (NWP)—have issued the *Framework for Success in Postsecondary Writing* (Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project , 2011) and the *WPA Outcomes Statement for First-Year Composition (v3.0)* (Council of Writing Program Administrators, 2014). The *Framework* identifies the rhetorical and 21st-century skills, as well as habits of mind and experiences, that are foundational for college success. Two pedagogies are recommended: writing, reading, and critical thinking experiences, with particular attention to rhetorical knowledge, critical thinking, writing processes, knowledge of conventions, and multimodal composing; and habits of mind that determine the ways that students approach learning, including curiosity, openness, engagement, creativity, persistence, responsibility, flexibility, and metacognition. The *Outcomes Statement* identifies knowledge and skills students should posses at the end of their first-year composition course in the areas of rhetorical knowledge, critical thinking, reading and composing, processes, and knowledge of conventions. When taken together, the panel and expert guidelines result in a defined construct that may be described as follows: Writing is a process involving modeling, practice, and reflection. Seen through a language arts lens, writing and reading should be taught concurrently to emphasize key writing features; seen through an assessment perspective, frequent and timely evaluation of student writing enhances student opportunity to learn (Moss et al., 2008). In terms of the cognitive domain, writing is best understood as an activity of cognitive complexity; regarding the intrapersonal domain, the act of writing is accompanied by known personality factors. Such a definition exists in a broad nomothetic span that, in turn, is drawn down through idiographic representation in unique assessment situations (Mislevy & Elliot, 2020)

In the 21st century, whether students write in standardized or classroom contexts, they are expected to have knowledge of writing based on pedagogies similar to those identified above.

However, just as little is known about the ways that inferences derived from standardized writing align with inferences drawn from coursework in terms of writing features, less is known about how models used to drive explicit instruction result in writing features that students use across contexts.

1.4. Automated Writing Evaluation

Automated writing evaluation (AWE) offers a novel opportunity to compare how students' standardized assessment and coursework writing represent their writing skills as understood through a defined construct model. AWE uses natural language processing (NLP) methods to capture a wide range of construct-relevant writing features, including argumentation (e.g., use of claims and sources), coherence (e.g., topic flow, use of transition terms), sentence structure (e.g., prepositional phrases), and knowledge of English conventions (e.g., grammar, usage, and mechanics; Burstein et al., 2018; Cahill & Evanini, 2020; Matsumura et al., 2020; Shermis & Burstein, 2013). Utility value measures have been developed to address the intrapersonal domain (Beigman Klebanov, Burstein, et al., 2017). The technology has been widely used in educational assessment and instruction typically as a means to generate feature measures that capture various aspects of an essay. These measures can then be used to either compute an overall score for the essay (e.g., in a high-stakes writing assessment) or to provide feedback on specific aspects of writing, such as in instructional applications (Foltz & Rosenstein, 2017). The present study is concerned only with providing feedback on specific aspects of writing. While we make no claim that these features fully represent the writing construct as defined above, we do claim that these targeted features provide a principled way to understand student writing and to provide structured automated feedback in real time to students.

The fine-grained construct-relevant feature measures generated by AWE systems allow for detailed performance feedback to student writers. There are several systems that use AWE to provide students with feedback about their writing; examples include ETS's *Criterion*® online essay evaluation (Burstein et al., 2004), Writing Mentor® (Burstein et al., 2018) systems, Grammarly®, Pearson's Write-to-Learn (Foltz & Rosenstein, 2017), Turnitin's Revision Assistant (Woods et al., 2017), DocuScope (Kaufer et al., 2006), and Writing Pal (W-Pal; Roscoe & McNamara, 2013). A comprehensive review conducted by Stevenson and Phakiti (2013) on the effects of using AWE feedback in writing classrooms concluded a modest positive effect of AWE use on students' writing quality. Wilson and Czik (2016) conducted an empirical study to evaluate the effectiveness of combining feedback generated by an AWE system and teacher response. The authors reported that this combination of feedback was associated with greater student persistence in writing. Similar findings were reported in other studies such as Grimes and Warschauer (2010), Cassidy et al. (2016), and Foltz (2014), demonstrating that the use of AWE feedback improved students' motivation during writing.

It is worth noting that the use of AWE for classroom-based teaching and learning is not without debate (e.g., Chen & Cheng, 2008; Hoang & Kunnan, 2016). Deane (2013) has

addressed the limits of automated feature analysis through a construct modeling approach in which "analysis of linguistic features of individual samples, combined with a rich array of electronically collected information, might enable us to identify appropriate instructional interventions without removing the writing from a channel in which it was written to humans for real communicative purposes" (p. 21). It is in this spirit of complementarity between automated feature analysis and teacher response, as proposed by Deane (2013, demonstrated in Hazelton et al. (2021), that the feature descriptions shown in Table 1 are provided.

Table 1

36 AWE Feature Descriptions

Argument feature measures (ARG Features 1-2) that quantify the following: (1) average number of claims (2) average number of claim verbs from an extended discourse cue lexicon from Burstein et al. (1998)
Organization and Development feature measures (OD Features 1-7) that quantify the following: (1) discourse text segments most likely in argumentative writing (such as thesis statements, main points, supporting details, and conclusion statements; Attali & Burstein, 2006) (2) length of discourse text segments most likely in argumentative writing (such as thesis statements, main points, supporting details, and conclusion statements; Attali & Burstein, 2006) (3) sentences associated with an argument (Beigman Klebanov, Gyawali, & Song, 2017) (4) distribution of topical keywords (Beigman Klebanov & Flor, 2013) (5) discourse quality scores (Somasundaran et al., 2014) (6) keywords associated with the largest topic (Beigman Klebanov & Flor, 2013; Burstein et al., 2016) (7) pairs of words in the text that are strongly semantically related based on the pointwise-mutual information measure (Beigman Klebanov & Flor, 2013; Burstein et al., 2016)
Sentence Structure feature measures (SSTR Features 1-7) (Madnani et al., 2016) that measure normalized counts of the following: (1) longer prepositional phrases containing at least two adjacent prepositional phrases (e.g., The cat sat *in the box on the table*.) (2) longer sentences containing one independent clause and at least one dependent clause (3) complex verbs (e.g., *did leave*) (4) complex noun phrases that can be one of two kinds of structures (e.g., *highly-qualified teacher*, *the teacher of the year*) (5) relative clauses and the noun referent for their pronoun (6) passive sentences. (7) sentence variety (Deane et al., 2021)

Vocabulary features (VCB Features 1-10) that quantify the following: (1) average word length (Attali & Burstein, 2006) (2) number of terms that belong to homonym sets (e.g., *to*, *too*, *two*; Burstein et al., 2004) (3) number of inflected word forms (Burstein et al., 2017) (4) number of derivational word forms (Burstein et al., 2017) (5) number of pronouns (6) number of stative verbs (i.e., express states vs. action, e.g., *feel*; Burstein et al., 2017) (7) presence of positive and negative sentiment vocabulary terms from the VADER corpus[2] (Burstein et al., 2017) (8) vocabulary richness, using an aggregate feature composed of a number of text-based vocabulary-related measures (e.g., morphological complexity, relatedness of words in a text; Burstein et al., 2017; Deane et al., 2021) (9) verbs used in metaphorical contexts (Beigman Klebanov et al., 2015; Beigman Klebanov et al., 2016) (10) aggregate measure generated related to word frequency (Attali & Burstein, 2006)
English Conventions features (CNV Features 1-7) that quantify different aspects of English conventions based on: (1) normalized, aggregate measure of grammar error counts (Attali & Burstein, 2006) (2) normalized, aggregate measure of mechanics error counts (Attali & Burstein, 2006) (3) normalized, aggregate measure of word usage error counts (Attali & Burstein 2006) (4) aggregate proxy measure for overall errors in grammar usage and mechanics (5) presence of contractions (Burstein et al., 2018) (6) aggregate measure related to collocation and preposition use (Burstein et al., 2013) (7) words and expressions related to a set of 13 "unnecessary" words and terms (such as *very*, *literally*, *a total of* ; Burstein et al., 2018)
Utility-Value Language feature measures (UVL Features 1-3) based on Beigman Klebanov et al. (2017) that count the instances of language that writers can use in personal reflections of the value of content to their personal lives. These measures quantify the following: (1) use of argumentative connectives (e.g., *furthermore*) and narrative elements (e.g., past tense verbs), which are often used in personal stories (2) everyday vocabulary related to the extent and specificity of personal reflection that connects course material to the writer's personal life (e.g., family) (3) use of grammatical categories that express reference to self-address and other people using first-person singular (e.g., *I*, *mine*) and plural pronouns (e.g., *we*, *ourselves*), and second-person pronouns (e.g., *you*), possessive determiners (e.g., *their*), and indefinite pronouns (e.g., *anyone*)

Notable in Table 1 is that the features are designed to reflect the results of pedagogical best practices established in both *Teaching Secondary Students to Write Effectively* (Graham et al., 2016) and the *Framework for Success in Postsecondary Writing* (CWPA, NCTE, and NWP, 2011). Specifically, the features reflect a targeted, sociocognitive approach to writing that incorporates attention to rhetorical knowledge (social, across-person practices) as manifested in targeted linguistic features (cognitive, within-person practices; Mislevy, 2018). In these features, attention is also paid to habits of mind, specifically metacognition, in the case of utility value. While writing processes can be captured and analyzed using keystroke logs (Guo et al., 2018), that work is beyond the scope of this study. Yet, as we discuss below, the features shown in Table 1 can be used to provide information related to the model-practice-reflect instructional cycle and the use of feature-based assessments of student writing to inform instruction and

[2] https://github.com/cjhutto/vaderSentiment

feedback. The model is described with discourse mode as the first (and most important) feature, with knowledge of conventions identified last among the rhetorical features. The sole intrapersonal feature identified in the present study is utility value, explained most fully in section 5.2.

Before leaving Table 1, we note that the features targeted in this study are similar to features targeted in a wide range of corpus studies. As Ringler et al. (2018) have noted, corpus analysis tools such as those used in the present study offer one promising set of methods by which to address task design and evaluation issues. As Ringler et al. note, corpus analytic methods have been useful in identifying patterns or types of language that recur across successful student texts, including register features (Brown & Aull, 2017), rhetorical moves (Cotos et al., 2015), and lexicogrammatical features (Hardy & Römer, 2013). In addition, Aull (2017) recently demonstrated how identifying recurring types of language can be useful for explicating genres of writing tasks and raising questions about task design. To provide interpretation and use arguments across corpus analysis studies, Ringler et al. (2018) have developed a three step process using sequencing, comparison, and diagnosis to provide evidence-based heuristics that can, in turn, be used to understand how specific writing tasks fit into a classroom pedagogical sequence as well as compare to larger genres of writing outside of their immediate, local writing classroom environment. These interpretation and use arguments are critical if we are to understand the relevance of corpus analysis studies across individual studies and institutional sites. Put simply, AWE must be understood within specific learning contexts—and the present study demonstrates the significance of situated learning.

2.0 Literature Review

Limited studies have been conducted that use AWE to examine postsecondary writing skills, as well as relationships between writing and broader academic performance outcomes—those variables that are operationally distinct from AWE and used to examine "criterion evidence" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Further, to our knowledge, few studies have been conducted that use AWE analysis to compare writing from assessment and coursework contexts and examine relationships between these two contexts and academic performance factors.

Beigman Klebanov et al. (2016) examined a sample of responses to the GRE argumentative essay writing assessment in comparison to a sample of undergraduate argumentative coursework writing samples. They note that the argumentative writing genre demonstrates stable characteristics across assessment and instructional contexts with regard to systematic linguistic characteristics, such as consistency of rhetorical choices. The article also shows differences. For example, the coursework writing included more features characteristic of informative writing as compared to the text of GRE argumentative writing from an independent sample. Perin and Lauterbach (2018) examined writing skills of low-skilled adults attending college

developmental education courses by determining whether variables from an automated scoring system were predictive of writing quality. The authors applied the Coh-Metrix system—an AWE approach that analyzes texts and generates features related to language and discourse, such as word concreteness, syntax, and cohesion (Graesser et al., 2004)—to students' written coursework to study whether Coh-Metrix features predicted writing skills as measured by human holistic rating of writing quality. They identified 10 Coh-Metrix variables that were predictive of writing skills. Beigman Klebanov et al. (2017) used AWE to extract features associated with *utility value* in student writing samples collected from an intervention study in an undergraduate STEM setting (Harackiewicz et al., 2016). Utility value in the writing samples expressed how a biology topic related to a student's life. The authors showed that the presence of words identified as being associated with personal reflection (e.g., *our*, *we*, *people*, *human*, *am*, *me*, *parents*, *brother*, *family*, *children*) was predictive of writing responses with higher utility-value scores assigned by human raters. The automated assessment of utility value may be important because the Harackiewicz et al. (2016) study showed that, in the context of the study, higher utility-value scores in students' writing were correlated with course success and progression to a higher-level biology course.

Using data from 108 university students, Allen et al. (2016) investigated how linguistic properties in students' writing can be used to model individual differences in postsecondary students' vocabulary knowledge and comprehension skills. Linguistic essay features were computed using the ReaderBench framework on responses to a standardized writing assessment task. The ReaderBench framework is an automated text analysis tool that calculates linguistic and rhetorical text indices on argumentative essays in response to a standardized constructed response test task. The authors tested for the relationship between these automated feature scores and students' Gates-MacGinitie Vocabulary and Reading Comprehension test scores and found that five features accounted for 45.3 percent of the variance in vocabulary scores and three accounted for 36.3 percent of the variance in comprehension scores. In a set of papers from a study of university students, Burstein and colleagues examined relationships between writing achievement as measured by selected AWE feature measures (Burstein et al., 2017; Burstein et al., 2019; Burstein et al., 2020; Ling et al., 2021). These studies use both standardized writing assessment and coursework writing samples and AWE and academic performance measures that are similar to features and performance measures reported in the present study. Across the studies, varying subsets of features were found to be related to performance measures. Vocabulary features were identified as statistically significant predictors for many of these measures. Findings from studies such as these contribute to an increased understanding of how AWE can provide insights about linguistic characteristics of students' writing in response to standardized assessments and coursework assignments, and how linguistic characteristics, as captured by AWE, relate to academic performance factors.

3.0 Research Questions

In this study, we examine the feasibility of using AWE to examine writing samples from standardized assessment and coursework contexts to assess how they compare and how they are related to measures of students' academic performance. The study addresses the following research questions (RQ):

- RQ1: What can AWE tell us about variation in student writing samples drawn from two specific performance contexts: standardized argumentative writing versus coursework argumentative writing?
- RQ2: What are the relationships between writing subconstructs (as captured by AWE) and other measures of students' academic performance?

4.0 Study Methods

4.1 Sites

As part of a larger study, six four-year universities participated in the original data collection. Included in the current analyses are writing samples submitted from five of the original sites (designated as A-E), shown in Table 2.

Table 2

Student Demographic & Site Composition

	N	%
Gender		
Female	117	65%
Male	63	35%
Race		
Black	92	51%
White	49	27%
Hispanic	23	13%
Asian	5	3%
Multi-race	3	2%
Native Hawaiian or other	1	1%
Pacific Islander		
Unreported	7	4%
Site		
A	86	48%
B	59	33%
C	28	16%
D	4	2%
E	3	2%

Note. Percents are rounded to the nearest integer.

Data from one of the six sites was excluded because it did not meet inclusion criteria. Specifically, no students from that site both completed the assessment and submitted argumentative coursework writing.

4.2 Participants

In the larger study, all participating students (*N*=735) were invited to take a standardized writing assessment and submit coursework writing to a web-based data-collection portal developed for the project. A subsample of 180 students from the larger study is included for this study. Table 2 provides the demographic breakdown for these students. This breakdown follows the U.S. Integrated Postsecondary Education Data System reporting categories. The 180 students constitute all study participants who both completed the standardized writing assessment and submitted at least one argumentative coursework writing sample. Students included in this analysis were exclusively enrolled in first-year English composition courses. The sample sizes from each site (A-E) are shown in Table 2. The majority of the sample students were enrolled in the two sites with the highest proportion of Black students, one of which was in the Historically Black College & University network (Site A). Consequently, Black students constitute 51 percent of the sample. The remaining students are mostly White (27%) or Hispanic (13%). Note that Site C was a Hispanic-serving institution. In terms of gender, the majority of the sample—65 percent—is female (as a category assigned, not as an identity preferred; see Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys, 2016).

4.3 Instruments

4.3.1 Standardized Test Writing Tasks

Participants were administered two HEIghten standardized test instruments developed by Educational Testing Service (ETS): the HEIghten Written Communication test and the HEIghten Critical Thinking test. The HEIghten Written Communication (WC) test measures four dimensions of a test-taker's written communication: knowledge of social and rhetorical situations, knowledge of conceptual strategies, knowledge of language use and conventions, and knowledge of the writing process. The HEIghten WC test was administered in a 45-minute testing session. In the first section, students were asked to complete an argumentative essay task to support their position on a topic with reasons and examples. The second section consisted of 24 passage-based multiple-choice items. Each participating student completed one of two WC test forms. The scores for the direct writing measure (i.e., essay) and multiple-choice portions are separate. The reported scale score for the multiple-choice section can range from 150 to 180; the scale score for the essay writing section ranges from 0 to 12. The HEIghten Critical Thinking test measures a student's analytic and synthetic skills. The test-taking time was no more than an hour and included 32 items. The reported scale score ranges from 150 to 180. The relationship

between critical thinking and argumentation is well documented and researched as two overlapping constructs (e.g., Godden & Walton, 2007; Liu et al., 2014). Furthermore, critical thinking is a key skill associated with academic accomplishment, learning transfer, and workplace success.

4.3.2 Coursework Writing

Students submitted writing samples from their coursework writing assignments. As mentioned earlier, data collection from the larger study attempted to collect all writing assignment data from participating courses. For this analysis, we identified only writing samples that are from the argumentative writing genre. To do this, the study research team developed a coding scheme to analyze and manually label instructor writing assignments according to three genres prominently discovered in the assignments: (1) argumentative, (2) informational, and (3) reflective. To ensure that the coursework writing data was of a similar genre as the writing samples from the standardized writing assessment instrument, the coursework assignment data included in this analysis were selected only if they were labeled as argumentative genre. (Note that the coursework writing data, student metadata, and coursework writing genre labels are publicly available. The standardized test data include undisclosed writing tasks, so these have not been publicly released. The data include 20 essays not included in the study because these essays diverged in some way from a standard argumentative essay.)

4.3.3 Institutional Data

Several key academic measures were collected for each participating student from the institutional databases at each site. These measures included SAT total scores, SAT Reading scores, ACT total scores, high school GPAs, and college GPAs. For students whose data contained only ACT scores, we converted those scores to the SAT scale (ACT, n.d.). We recognize that scores from these assessments do not serve all students equally; yet these standardized measurements do allow a sense of evidence related to criterion measures (Sireci & Randall, 2021; Sireci & Talento-Miller, 2006).

4.4 AWE Feature Generation

For all texts from the HEIghten WC assessment and the coursework, AWE tools were used to generate the 36 writing features shown in Table 1. These features represent six writing subconstructs: Argumentation (2 features, e.g., claim terms), Organization and Development (7 features, e.g., text coherence), Sentence Structure (7 features, e.g., use of clauses), Vocabulary (10 features, e.g., word complexity), English Conventions (7 features, e.g., grammar), and Utility-Value language (3 features, i.e., words expressing connections between current learning and individual experiences). The mapping of the features to the subconstructs was heuristically determined by the feature developers on the basis of the structure and intended measurement of the AWE features associated with the defined writing construct. The next

section describes the steps taken to create the six composite feature measures (scores) that used the set of 36 features.

4.5 Data Analysis

For each subconstruct feature, the corresponding features (listed in Table 1) were aggregated to create a subconstruct score. There are two main reasons to use aggregated subconstruct scores instead of single features. One reason is that composite scores have higher internal consistency than single features when the features are theoretically related to the same (sub)construct. This is similar to the total test score being more reliable than any single item score. The second reason is that it is more manageable to evaluate and make interpretations of six measures compared to 36. To determine the weights of those features in a subconstruct, principal component analysis (PCA) was conducted separately for each subconstruct using the relevant features. The choice of PCA was made to find an aggregation rule that can maximize the reliable variance explainable by the features. It was not conducted to test the structure of the features or test the validity of the subconstruct scores. The interpretation of the subconstruct scores is supported by content of the features and the expert judgment used in creating their groupings. In order to compare the HEIghten and coursework writing samples, we need a common set of feature weights for the writing samples from both sources. To ensure that the features from the two data sources are on a common scale, each AWE feature was first standardized across the data sets. Specifically, the original feature values were divided by this quantity $\sqrt{(SD_H^2 + SD_U^2)/2}$, where SD_H and SD_U are the feature standard deviations in HEIghten and coursework data sets, respectively. Then, each feature was centered to have a mean of 0 within each data source in order to remove the source effect. The PCA was then conducted on the combined sample of HEIghten and coursework responses (n = 360). The feature loadings in the first principal component were used as feature weights for each subconstruct. Finally, the weights were applied to the standardized feature values (before centering) in each data set to generate six subconstruct (or composite) feature scores. (For more on establishing weights of linguistic units, see Stefanowitsch [2020, pp. 90-93].) All the subsequent data analyses were conducted on the composite scores.

To answer RQ1, we used different statistical analysis and data visualization methods. The distributions of the six composite feature scores between the HEIghten and coursework responses were compared in histograms and boxplots. Paired t tests were conducted for each composite feature to examine whether the means of the composite scores were statistically the same (H0). To answer RQ2, we used various academic outcome variables including high school GPA, college GPA, SAT total scores, SAT Reading scores, and HEIghten Critical Thinking scores. We also examined the correlations of composite feature scores with the scores students received from the multiple-choice portion of the HEIghten WC (Written Communication) test.

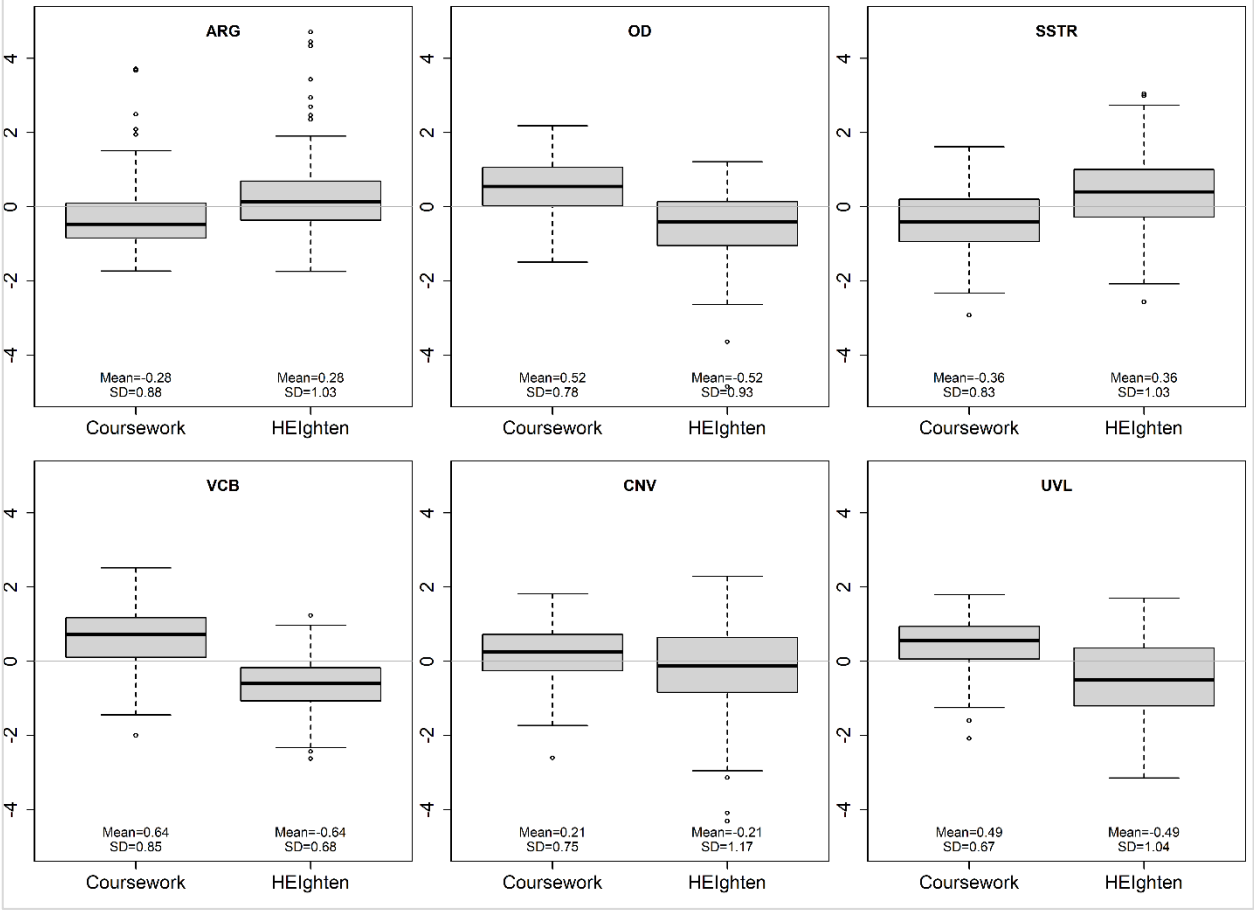
5.0 Results

5.1 Feature Distributions Across Two Contexts

Figure 1 compares the boxplots of the six composite feature scores between HEIghten and coursework writing samples. Each panel includes the mean and standard deviation for composite scores for HEIghten and coursework samples. Because the composite scores were scaled to have mean zero for the combined sample of HEIghten and coursework data, means for the two groups must sum to zero. The boxplots, means, and standard deviations reveal notable differences in the distributions of scores from the two performance contexts for nearly all of the composite subconstructs. Table 3 presents the *t* test for differences in the means; all were significant at the 0.001 level.

Figure 1

Comparison of Composite Feature Scores: Boxplots, Means, and Standard Deviations (SD)



Note. ARG: Argumentation; OD: Organization and Development; SSTR: Sentence Structure; VCB: Vocabulary; CNV: Convention; UVL: Utility-Value Language.

Notably, the HEIghten responses demonstrated significantly higher composite scores for Argumentation (ARG) and Sentence Structure (SSTR) as compared to the coursework writing: $t(179) = 5.6$, $p < 0.0001$ for ARG and $t(179) = 7.99$, $p < 0.0001$ for SSTR. By contrast, for the remaining and complementary four composites, Organization and Development (OD), Vocabulary (VCB), Conventions (CNV), and Utility-Value Language (UVL), the coursework samples showed higher values than HEIghten responses: $t(179) = -11.81$ for OD; $t(179) = -19.13$ for VCB; $t(179) = -4.38$ for CNV; and $t(179) = -10.34$ for UVL. All p-values were smaller than 0.001.

Table 3

HEIghten and Coursework Composite Score Correlations

	***t* tests for mean differences (HEIghten minus coursework)**		**Correlation of subconstruct composite scores between HEIghten and coursework**		**Correlation of subconstruct composite scores between 2 coursework samples**	
Subconstruct	***t*-statistic**	***p*-value**	***r***	***p*-value**	***r***	***p*-value**
ARG	**5.60**	< 0.001	0.04	0.60	-0.06	0.56
OD	**-11.80**	< 0.001	0.06	0.41	**0.42**	< 0.001
SSTR	**7.99**	< 0.001	**0.17**	0.02	**0.20**	0.04
VCB	**-19.13**	< 0.001	**0.32**	< 0.001	**0.47**	< 0.001
CNV	**-4.38**	< 0.001	0.14	0.06	0.09	0.37
UVL	**-10.34**	< 0.001	-0.04	0.55	**0.35**	< 0.001
Sample size	180		180		108	

Note. ARG: Argumentation; OD: Organization and Development; SSTR: Sentence Structure; VCB: Vocabulary; CNV: Convention; UVL: Utility-Value Language. Bolded values are statistically significant at $p < 0.05$ level.

Table 3 also contains the Pearson correlations between the composite features scores for each subconstruct for HEIghten and the coursework. The values indicate that the composite feature values for writing from the two different contexts are weakly related with absolute correlation values ranging from 0.04 to 0.32. The only two significant correlations are $r = 0.17$ for SSTR and $r = 0.32$ for VCB between the two contexts. If the composite feature values from the two contexts create similar relative orderings of students, these correlations would be close to 1.0. Instead they are small and statistically significantly greater than zero only for SSTR and VCB. It is, however, worth noting that the features from each context are based on a single writing sample (i.e., one response to a single standardized writing task or a single response to a coursework assignment). Values could be sensitive to the topic, which could depress the correlations across contexts (or with other factors, such as GPA). Students only completed a single standardized task, but a subset of the 180 students (i.e., $n = 108$) uploaded two

argumentative essays in their coursework. For these students, the correlations on the six composites between the two coursework samples are also given in Table 3. The correlations are low to moderate (ranging from absolute values 0.06 to 0.47), indicating considerable sensitivity of the subdomain composite scores to the particular writing sample, which could suppress the correlation of the scores from coursework writing with HEIghten writing. One other notable result is that, for OD and UVL, the correlations between two coursework samples are considerably larger than with the values from the HEIghten responses (i.e., 0.42 vs. 0.06 for OD, 0.35 vs. -0.04 for UVL). This result further suggests that writing in the two different contexts is likely to provide information that would tend to consistently rank students differently for some dimensions of writing.

Table 4

*Inter-Indicator Correlations for HEIghten and Coursework Writing (*N = *180)*

	ARG		OD		SSTR		VCB		CNV		UVL		Length	
	H	**C**	**H**	**C**	**H**	**C**	**H**	**C**	**H**	**C**	**H**	**C**	**H**	**C**
ARG	**1.00**	**1.00**	-0.15*	-0.15*	0.20**	0.57**	-0.25**	-0.42**	0.11	0.21**	0.04	-0.14	-0.09	-0.19*
OD	-0.15*	-0.15*	**1.00**	**1.00**	0.04	-0.18*	0.20**	0.42**	0.23**	-0.17*	0.34**	0.06	0.53**	0.61**
SSTR	0.20**	0.57**	0.04	-0.18*	**1.00**	**1.00**	-0.10	-0.49**	0.18*	0.25**	0.18*	0.20**	0.05	-0.10
VCB	-0.25**	-0.42**	0.20**	0.42**	-0.10	-0.49**	**1.00**	**1.00**	-0.03	-0.28**	-0.15*	-0.39**	0.00	0.48**
CNV	0.11	0.21**	0.23**	-0.17*	0.18*	0.25**	-0.03	-0.28**	**1.00**	**1.00**	0.19*	0.01	0.25**	-0.25**
UVL	0.04	-0.14	0.34**	0.06	0.18*	0.20**	-0.15*	-0.39**	0.19*	0.01	**1.00**	**1.00**	0.72**	-0.10

Note. H: HEIghten responses; C: Coursework responses. * : $p < 0.05$; **: $p < 0.01$. ARG: Argumentation; OD: Organization and Development; SSTR: Sentence Structure; VCB: Vocabulary; CNV: Convention; UVL: Utility-Value Language.

Table 4 shows the correlations among the values from the six composite scores for each of the two contexts. The correlations are generally small in both writing samples. Consistent between the two writing contexts, VCB is negatively correlated with all of the other indicators with the exception of OD. SSTR is positively correlated with ARG, CNV, and UVL, but negatively correlated with VCB in both contexts and negatively correlated with OD for coursework essays. It is also noted that these correlations are greater in the Coursework writing compared to HEIghten writing. There are cases where the correlations differ between the two contexts. For instance, OD and CNV are positively correlated in the HEIghten data, but negatively correlated in the coursework writing. Furthermore, although VCB is correlated with most of the other composite scores in both contexts, the correlations are consistently stronger for coursework. Also, VCB is relatively strongly correlated with length in the coursework, but it is uncorrelated in HEIghten, suggesting that the VCB measure functions differently across contexts. We also see this pattern when we examine correlations with criterion variables below.

5.2 AWE Composite Scores and Performance Indicators

Table 5 contains the correlations of each of the feature scores for the HEIghten responses and coursework, along with the correlations with the performance measures. Table 6 contains the same analysis controlling for response length. The results suggest distinct differences between the contexts.

Table 5

Correlation with Criterion Variables

		HEIghten						Coursework					
Criterion variables	***N***	**ARG**	**OD**	**SSTR**	**VCB**	**CNV**	**UVL**	**ARG**	**OD**	**SSTR**	**VCB**	**CNV**	**UVL**
High school GPA	164	-0.02	0.29**	-0.03	0.27**	0.22**	0.09	0.06	0.03	-0.03	0.09	0.11	-0.25**
HEIghten WC Multiple Choice	177	-0.03	0.21**	-0.08	0.25**	0.18*	0.11	-0.01	-0.09	-0.14	0.07	0.17*	-0.23**
HEIghten Critical Thinking Score	168	-0.06	0.19	0.03	0.31**	0.27*	0.29**	-0.03	-0.11	0.02	0.18	0.20	-0.28*
Current Semester GPA	178	-0.02	0.22**	-0.08	0.27**	0.28**	0.16*	-0.01	-0.01	-0.17*	0.12	0.09	-0.16*
+1 Semester GPA	138	-0.08	0.09	-0.09	0.21**	0.19*	0.20*	0.00	0.16*	-0.06	0.12	-0.02	-0.15*
+2 Semester GPA	133	-0.05	0.20**	-0.05	0.23**	0.14	-0.03	-0.10	0.06	-0.13	0.17*	0.08	-0.22**
+3 Semester GPA	57	-0.04	0.30**	-0.03	0.29**	0.13	0.01	-0.10	0.09	-0.16*	0.18*	0.09	-0.23**
SAT Total Score	161	-0.02	0.27**	-0.01	0.26**	0.17*	0.04	-0.04	0.15	-0.11	0.19*	0.08	-0.21*
SAT Reading Score	93	-0.07	0.19	0.16	0.34**	0.41**	-0.04	-0.12	0.32*	-0.06	0.20	0.13	-0.12

Note. * : $p < 0.05$; **: $p < 0.01$. ARG: Argumentation; OD: Organization and Development; SSTR: Sentence Structure; VCB: Vocabulary; CNV: Convention; UVL: Utility Value. +1, +2 or +3 Semester GPA: GPA collected 1, 2, or 3 semesters after the students enrolled in the study.

Table 6

Correlation with Criterion Variables Controlling for Response Length

		HEIghten						Coursework					
Criterion variables	***N***	**ARG**	**OD**	**SSTR**	**VCB**	**CNV**	**UVL**	**ARG**	**OD**	**SSTR**	**VCB**	**CNV**	**UVL**
High school GPA	164	0.01	0.23**	-0.04	0.27**	0.18*	-0.07	0.07	-0.04	-0.02	0.05	0.14	-0.24 **
HEIghten WC Multiple Choice	177	0.002	0.07	-0.10	0.29**	0.22**	-0.08	-0.03	-0.05	-0.15	0.13	0.16*	-0.23**
HEIghten Critical Thinking Score	168	-0.05	-0.07	-0.10	0.22**	0.14	0.01	-0.02	-0.17	0.03	0.18	0.22*	-0.28**
Current Semester GPA	178	-0.04	0.19*	-0.05	0.23**	0.13	-0.12	-0.00	-0.04	-0.16*	0.11	0.11	-0.16*
+1 Semester GPA	138	-0.02	0.27**	-0.04	0.29**	0.09	-0.14	0.02	0.11	-0.05	0.07	0.01	-0.14
+2 Semester GPA	133	-0.01	0.22*	-0.01	0.26**	0.13	-0.11	-0.08	0.00	-0.12	0.15*	0.11	-0.21**
+3 Semester GPA	57	-0.06	0.18	0.15	0.34**	0.41**	-0.12	-0.08	0.03	-0.15	0.15	0.12	-0.22**
SAT Total Score	161	-0.01	0.13	-0.09	0.25**	0.14	-0.05	-0.01	0.08	-0.10	0.14	0.12	-0.21*
SAT Reading Score	93	-0.05	-0.01	0.01	0.32**	0.17	0.05	-0.09	0.21	-0.02	0.11	0.20	-0.10

Note. * : $p < 0.05$; **: $p < 0.01$. ARG: Argumentation; OD: Organization and Development; SSTR: Sentence Structure; VCB: Vocabulary; CNV: Convention; UVL: Utility Value. +1, +2 or +3 Semester GPA: GPA collected 1, 2, or 3 semesters after the students enrolled in the study.

For the HEIghten responses, OD, VCB, and CNV are positively correlated with multiple external measures. These AWE feature measures are scaled so that higher values would tend to indicate higher quality writing. As shown in Table 6, UVL is also positively correlated with the scores on the HEIghten tests, SAT Reading, and HEIghten critical thinking scores. However, when we control for the length of the essay, the correlation between UVL and the external indicators is notably smaller and no longer significant. This suggests that the relationship was spurious and due to the fact that students who write longer essays for HEIghten tend to have higher values on the indicators rather than the UVL actually providing uniquely informative data. The significant correlations of CNV and OD with external measures generally become weaker in strength after controlling for response length, and half become insignificant. The correlations of VCB with external measures all remained significant after controlling for length, with limited or no degradation in strength. For the coursework writing data, the UVL is consistently and negatively correlated with the external outcome indicators, even if we control for length. VCB is correlated with GPA variables, but VCB is also related to the length of the writing sample. The only other significant, yet weak correlations are CNV with SAT scores, and SSTR with the scores on the writing multiple choice items from HEIghten and college GPA one semester after the active data collection semester (+1 semester GPA). Overall, there are limited significant correlations between other AWE composite scores and external measures in coursework responses, for which the pattern largely remained after controlling for response length.

6.0 Discussion

6.1 Study Findings: Overall Inferences

To our knowledge, this exploratory writing analytics study is unique in its comparison of standardized and coursework writing in this way: (1) using AWE features to study writer production of linguistic features associated with argumentative writing in the two contexts, and (2) exploring the relationships between linguistic feature measures in writing and academic performance factors. Overall, the study shows that differences exist in the feature distributions associated with writing in these different contexts. Findings suggest that student writing output from each of these contexts tells us part of the story about what a student can produce in a single piece of writing and consequently may provide different information about the students' writing skills. For instance, students may be less likely to produce well-formed text that is free of grammar and spelling errors on a standardized writing assessment due to time and resource constraints. By contrast, students may be more likely to prepare well-formed texts for a coursework assignment. Similarly, in standardized assessments, we may not see the breadth of students' vocabulary knowledge that we might find in their coursework writing when they have access to a thesaurus, or simply, just more time to wordsmith. These findings suggest that more and different types of writing are required to get a more complete picture of students' abilities for different writing subconstructs and to draw inferences about relationships between writing ability and academic performance.

The study focused on argumentative writing produced by students in response to a HEIghten standardized argumentative writing task and argumentative coursework writing tasks. The writing was evaluated on AWE composite feature dimension scores grouped into six subconstruct components of writing—shown in Table 1—for evaluating argumentative writing. Study findings reveal that AWE applied to writing samples collected from two different contexts—a standardized writing assessment and classroom coursework writing—may yield different inferences about a student's writing and writing ability.

Notably, the distributions of student AWE component scores differed between the standardized assessment and coursework writing. Therefore, any interpretation based on the magnitude of AWE dimension scores would need to be context-specific, and adjustments for context would be necessary. Similarly, using the data from the two contexts to make decisions about the relative abilities of different students on the six AWE components could lead to different inferences. Moreover, these differences are greater than differences from using different coursework samples for the same student, suggesting that the differences illustrated by our analysis are based on context. Such findings support the large number of calls for localization in writing assessment (Condon, 2013).

The relationship of scores for each of the six components and external factors further illustrates distinct differences. Most prominently, higher levels of UVL usage in argumentative coursework writing are associated with poor performance on nearly all of the external academic

performance indicators, while this is not the case for responses to the standardized HEIghten argumentative task. Alternatively, with higher scores on VCB (i.e., suggesting more sophisticated vocabulary usage) and on the measure of adherence to writing conventions as measured by CNV, responses to the HEIghten task are positively and significantly associated with all the indicators. By contrast, in coursework writing, VCB is only significantly associated with GPAs from the study semester (i.e., typically the student's first or second semester) and first and second semester GPAs after entering the study, while the CNV is only significantly associated with SAT total scores.

6.2 Study Findings: Close Reading of Utility-Value Language Use

Deeper consideration of these differences provides some insights into two performance contexts for writing and their potentially complementary contributions to understanding students' writing production in these contexts and relationships to academic performance indicators. Let's consider the case of utility-value language usage. The HEIghten tasks elicit students' evaluation of an argument based on brief background information provided as part of the writing item stimulus. The task instructions include text indicating that students can use their personal experiences to support their evaluation of the argument provided in their response. Thus, in the context of the standardized tasks used in this study, utility-value language usage is warranted; therefore, it is reasonable that it would not negatively affect writing quality. For coursework writing, however, utility-value language usage showed significant negative correlations with external measures of college GPA, suggesting that more use of utility-value language is associated with lower GPA (i.e., decreased likelihood of academic performance). However, looking at the correlations is not sufficient. In Table 7, we provide examples from the coursework writing to illustrate high and low utility-value language (UVL) usage.

Table 7

Excerpts from Coursework Writing with Higher and Lower UVL Values

High UVL 1	*"Amaka is an important character who helped influence Kambili. Amaka is the daughter of Aunty Ifeoma. In Amaka's house they're are religious but they sing prayers instead of saying it compared to Kambili household. Amaka doesn't eat at the dinner table in her house and doesn't have a set schedule which made Kambili surprised and puzzled. In the novel, Aunty Ifeoma encourages her kids to speak up for themselves which makes Kambili change in the story. Also, in Amaka's house they are more laughter and talking then Papa Eugene's house. "So, your voice can be this loud, Kambili," she said "(Adichie p170). Amaka is telling Kambili that she can talk loud in her household because she is not used to things like this. "Because you're bored with it? If only we all had satellite so every day could be bored with it" (Adichie p79)."*

High UVL 2	*"After speaking with Mrs. Nora Jones, Business professor at Community College I discovered some interesting things about the world of Accounting; payroll accounting specifically. On October 23, 2017 we discussed over the telephone and E-mail things about her previous job as a Payroll Technician. The following questions explain answers to my research into this field.* *What information that you learned from this experience particularly surprised or enlightened you?* *While I enjoy working with numbers, it can be frustrating in Payroll Accounting because if one is off by 0.10 cents then the error must be found in entire journal entries. This can take anywhere from five minutes to two hours. One must be patient when working as an Payroll Technician. Working as a payroll technician is like a puzzle, each individual employee has their own different things that are calculated into their pay such as: overtime pay, substitute pay, overload course hours and retirement pay."*
Low UVL	*"A social construct is an idea that is created by a society. What it means to that say race is socially constructed is that race and the different kinds of races are not based in biology, but are chosen, sometimes arbitrarily, by a society. An example would be how in the United States in the early 20th century when the law enforced segregation and discrimination, the courts decided who was racially white and who was not. The courts ended up concluding that Japanese were not white despite their light skin color, because pseudoscience of the time said so. However, the courts later concluded that Indians were not white, even though the same pseudoscience sources they used before indicated that Indians were white. This shows that definitions of race were defined not by shade of skin color or by the pseudoscience of the time, but were instead chosen arbitrarily based on preexisting prejudices against certain ethnic groups."*

In the High UVL 1 case, the writer discusses family relationships drawn from a novel. Language about oneself or related to family are typical utility-value terms evaluated in the UVL measure. The writer's vocabulary used to discuss the book is relevant to the novel, and use of UVL seems appropriate. However, other vocabulary in the excerpt is not particularly academic in contrast to the High UVL 2 and Low UVL. In High UVL2, language patterns such as "one," "her," and "their own" are indicative of utility value (i.e., discussion of how situations relate to oneself, others, or society) and appear appropriate to the topic. We also observe in High UVL 2 that the writer otherwise uses what would be considered academic language. As we might expect, in the Low UVL case, the topic does not appear to elicit UVL. The writing does not reflect personal experience and uses academic vocabulary that might be expected in undergraduate-level writing.

When using AWE to study finer-grained dimensions of writing beyond overall ability measured by a single summative score, the specific wording of the writing tasks may be important for determining how to interpret the AWE feature values. For utility-value language, a more nuanced set of features may be necessary to differentiate between the high-quality and problematic uses as we observed in the qualitative reviews of a small number of samples in Table 7. In a future analysis, a thorough, qualitative analysis of the coursework assignments might help to reveal ways to refine existing AWE feature measures, such that they are more informative about the quality of the students' writing for a given assignment. Such analysis

supports the findings of Melzer (2014) about the significance of the writing task, especially in classroom assignments in which limited informative purposes—with limited rhetorical situations that did not extend beyond the teacher as audience and examiner—are in evidence. In similar fashion, AWE study would benefit from more detailed knowledge about exploratory classroom writing tasks that invite students to make connections between their current learning and their own experiences.

For standardized writing assessments, like HEIghten, students do not have access to supports, such as grammar and spelling checkers and peer review. This is a criticism of such standardized assessments, and we are aware that such criticism of standardization is often related to concerns of equity and fairness (Cramer et al., 2018). Students are more likely to have such access, however, when writing in coursework contexts. Access to external writing support might create differences in the meaning of AWE feature scores. Related to the matter of writing support, students in this study were asked to complete a brief survey when uploading a coursework assignment. The survey consisted of five items including one that asked if anyone had helped them on the assignment (Response options were: a parent, brother, sister or other relative; a teacher; a tutor; someone else; no one helped me) and another item that asked if the student used a spell- or grammar-checker while writing. The large majority of students (86%) reported using spell- or grammar-checker. These aids could remove conventions problems that the CNV feature is designed to identify so that the feature values observed in coursework writing might not reflect students' unaided ability to follow such conventions. The CNV measure from the HEIghten would reflect only use of conventions assessed in the feature with no outside help. On-demand writing, such as in HEIghten context, might also put different demands on students' knowledge of vocabulary than would coursework writing. Specifically, in a standardized assessment, students cannot access any sources other than the task materials, whereas in coursework, students have more time and access to other resources to help in the selection of the vocabulary for the text. This classroom affordance might explain the differences we see in the relationship between Vocabulary and the external outcome measures. Both contexts might provide meaningful information about students' knowledge of vocabulary, but the interpretation needs to be context-specific. In addition, the AWE measure of vocabulary used in this study is a measure of the presence of sophistication of vocabulary usage, but not the proper usage of the vocabulary. Measures evaluating the task-appropriate usage of vocabulary might provide additional insights into students' writing skills, and this might behave differently in the two contexts. Vocabulary might also be specific to the task, and this would need to be further investigated.

An important lesson from the results for both HEIghten and the coursework writing is the importance of using feature values captured from multiple writing samples from each context to produce reliable measures of each writing subconstruct within each context. For both contexts, many of the correlations are small. One reason for these small correlations is the availability of only a single writing sample from each mode, and there is likely to be substantial variation across writing samples in the composite scores for each dimension. There are multiple reasons to

expect that scores for some dimensions will be sensitive to the task in standardized assessments; for instance, the vocabulary or utility-value language measures might depend on the language and instructions of the task. Similarly, the requirements of an assignment among other factors could affect dimension scores in coursework too. This variation was demonstrated for the coursework. We have two coursework samples for a subsample of students, and we found that the correlation between the scores for the same dimension from the two samples was low to moderate, indicating notable variation across the samples.

6.3 Limitations

The current study has four limitations that must be considered in the interpretation of the results: study sample, genres examined, standardized criterion measures, and shifting educational pedagogies.

As noted throughout, the study results may be sensitive to the small sample size in terms of number of students and number of institutions. The small sample reduces the statistical power of the tests of the relationships between feature values and other performance and outcome measures. Also, the results from this small sample of universities might not generalize to other institutions, especially highly selective four-year institutions or two-year institutions, since the sample includes only four-year institutions that were not highly selective. This sample is also somewhat unrepresentative of the four-year university student population in that the current sample has a greater proportion of Black students than the overall population. Additionally, we have no information on student English language proficiency status. The impact of the context and the functioning of the features could differ for English language learners and English proficient students. The sample is too small to explore how the results might vary across students of differing language backgrounds. Hence, although results clearly show the potential for assessments of writing to be sensitive to context, replication of the study would be useful to understand generalizability of the specific findings across the general population of university students.

In terms of genres examined, the study is also limited to argumentative writing for English composition courses. The coursework sample was restricted to assignments from this genre to match the genre of the HEIghten Written Communication assessment. Focusing on argumentative writing assignments may limit the writing expectations of the assignments and the resulting nature of the features. A broader corpus of postsecondary coursework and standardized assessment could support a deeper exploration of the impact of context on assessment of students' writing skills. Similarly, the 36 AWE features do not cover all aspects of writing. Other approaches to analyzing the writing might reveal other ways the context matters. For instance, Ringler et al. (2018) study the rhetorical strategies of argumentative writing beyond the 36 features used in the present study. As well, it is important to understand that the features themselves have limitations. For instance, the conventions feature, CNV, assesses adherence to traditional academic English conventions. A subset of the other features was also developed for

scoring standardized assessments of academic English. The utility of AWE to assess alternatives to academic English and how context might affect such assessments cannot be inferred from this current exploration. As demonstrated by our sample and the content of the current standardized tests, academic English remains of considerable focus in postsecondary education, and the current study provides useful considerations for assessing students' skills from that perspective. However, expanding AWE to alternative perspectives could be valuable in supporting more socioculturally responsive assessment of students' written communication abilities. As Gere et al. (2021) have noted,

> Standard language ideologies have embedded injustice in many of the policies and publications that guide writing instruction and assessment. These ideologies privilege certain forms of language as "correct," "better," or "commonsensical." Privileged forms are codified and enforced as "standard," while the language varieties and discursive patterns of less privileged groups receive discrimination and ridicule. (p. 385)

One potential alternative perspective is critical language awareness —pedagogies involving "exploration of language change, historical processes of standardization, distinctions between descriptive and prescriptive grammars, and an awareness that terms like 'conventions' are grounded in standard language ideologies" (Gere et al., 2021, p. 385). Further research involving the usefulness of critical language awareness to examine language ideologies might follow theory of action frameworks used by Hazelton et al. (2021). Specifically, future research might focus on the ways that social justice frameworks could be used to examine the ways that standard language ideologies are related to structural injustice at the level of knowledge of conventions, as well as the ways that these language ideologies are related to professional achievement and barriers to success (Gubala et al., 2020).

The use of standardized testing, commonly used as a form of criterion-related evidence as shown in the present study, has changed since our data has been collected, with many institutions going test optional for admissions (ACT, 2021). As well, recent research has demonstrated the ways that existing information on student performance might be used to determine course placement (Bahr et al., 2019). Two universities in our study were test optional as of 2022. The others had waived the test requirement in response to COVID but had not made tests optional. The changing landscape around admissions testing may change the relevance of standardized assessments and, as our study demonstrates, standardized assessment may provide an incomplete picture of students' ability. Given the various limitations listed above, it is risky to speculate how feature-based AWE will provide criterion-evidence when new methods are used for admissions, placement, and progression. That said, AWE features—as they presently exist and as they are modified by critical language awareness—may hold the potential to provide specific information on language use that is valuable in supporting educators in assessment decisions related to the advancement of opportunity (Moss et al., 2008).

While this study was conducted before the beginning of the global COVID-19 pandemic, there are implications for the use of feature-based AWE in classroom settings. A report by the U.S. Department of Education (2021) has documented

> the many ways that COVID-19, with all of its tragic impacts on individuals, families, and communities, appears to be deepening divides in educational opportunity across our nation's classrooms and campuses. Although the pandemic's effects will be studied for many years to come, we know from early studies that for many students, the educational gaps that existed before the pandemic—in access, opportunities, achievement, and outcomes—are widening. (ii)

To lessen the conditions under which disparate impact arises, the report identifies ensuring resource comparability across educational settings, providing English appropriate language support, ensuring inclusion, and incorporating academic adjustments and modifications. In achieving these objectives, feature-based AWE plays a unique role in that the concept was born digitally. Whether pedagogies are synchronous, asynchronous, or blended, AWE can provide important support in terms of writing feedback to students and learning analytics to administrators. Use of AWE-backed analytics might allow for flexibility in instruction and learning that might combat some inequities that the response to COVID identified, highlighted, or created. For example, feature-based AWE could improve self-directed writing instruction along the lines of Burstein et al. (2018), which might help students with inflexible schedules due to other responsibilities (e.g., childcare).

7.0 Conclusion

The community of writing analytics researchers and teachers is likely to continue to enhance AWE systems. As they do this, they should consider insights from studies such as this one to inform system enhancements and potential uses. Important in this work is to establish relationships between writing analytics (defined by Shum et al. [2016] as "the measurement and analysis of written texts for the purpose of understanding writing processes and products, in their educational contexts, and improving the teaching and learning of writing" [p. 481]) and learning analytics, especially the capacity of the field

> (1) to estimate the probability of student success in a course prior to and during the offering of a course; (2) to identify points at which instructional interventions might increase the likelihood of student success in a course; and (3) to support retrospective analysis of instructional materials, instructional interventions, and instructor teaching effectiveness. (Palmquist, 2019, p. 2)

Following Shum and Crick (2016), we note that the goal of establishing relationships between AWE systems and learning analytics is

> to forge new links from the body of educational/learning sciences research—which typically clarifies *the nature of the phenomena* under question using representations and language for *researchers*—to documenting how *data, algorithms, code, and user interfaces come together through coherent design* in order to automate such analyses—providing actionable insight for the *educators, students, and other stakeholders* who constitute the learning system in question. (p. 8)

AWE system developers would greatly benefit from identifying resonances with learning analytics when designing system enhancements to ensure that the innovation will support students in receiving immediate and personalized feedback relevant to a breadth of writing assignment types in varying contexts. For instance, if a student needs to prepare for a timed assessment, perhaps having access to an AWE practice environment including a timer and feedback relevant to improvement *after* the writing time is "up" could support students in on-demand test prep scenarios. For coursework, on the other hand, having immediate feedback for a breadth of AWE feedback types relevant to an assignment, and the ability to continuously revise and refine, is important in advancing opportunity to learn. In all contexts, the role of the instructor is fundamental in providing instruction that situates the role of AWE among various forms of feedback (such as peer review) and various pedagogies (such as critical language awareness). Study findings related to the relationships between AWE feature component measures and academic performance factors such as those found in the present study provide insights suggesting that AWE integration into varied performance contexts may prove useful to students and their instructors.

Note

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305A160115 to the Educational Testing Service. The opinions, findings, conclusions, or recommendations expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Acknowledgments

We are grateful to our partner institutions who made this work possible: Bloomsburg University of Pennsylvania, Bowie State University, California State University-Fresno, Jacksonville State University, Slippery Rock University, and University of North Carolina-Wilmington. We also thank Norbert Elliot for his invaluable guidance on the study design and insightful comments that shaped the paper in its current form.

Author Biographies

Dan McCaffrey is an Associate Vice President for Psychometric Analysis and Research at Educational Testing Service (ETS), overseeing psychometric and data analysis for ETS's testing programs and contracts such as the National Assessment of Educational Progress, the Programme for International Student Assessment, and state elementary and secondary testing. His research interests include machine scoring of constructed responses, measurement of student achievement growth, and causal modeling.

Mo Zhang is a Senior Research Scientist at Educational Testing Service. Her research interests include developing performance-based assessment, human and machine scoring of constructed-response items, modeling test-response process, and using educational data mining to generate learning analytics and performance feedback.

Jill Burstein is a Principal Assessment Scientist at Duolingo, Inc. and leads and conducts assessment research for the Duolingo English Test. Jill's research focus has been artificial intelligence (AI) in education. Prior to joining Duolingo, she led natural language processing (NLP) teams that designed innovative automated writing evaluation systems used in large-scale high-stakes assessment and digital writing support applications.

References

ACT. (2021). *Summary findings: Survey of higher education enrollment and admissions officers*. https://chronicle.brightspotcdn.com/a1/52/dcf530674ab59217fc5f025d6a76/210212-hedsurveysummaryfindings-externaluse.pdf

ACT. (n.d.). *2018 ACT/SAT concordance tables*. https://www.act.org/content/dam/act/unsecured/documents/ACT-SAT-Concordance-Tables.pdf

Allan, E. G., & Driscoll, D. L. (2014). The three-fold benefit of reflective writing: Improving program assessment, student learning, and faculty professional development. *Assessing Writing*, *21*, 37-55.

Allen, L., Dascalu, M., McNamara, D. S., Crossly, S., & Trausan-Matu, S. (2016). Modeling individual differences among writers using ReaderBench. In *Proceedings of the EDULEARN16 Conference* (pp. 5269-5279).

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Applebee, A., & Langer, J. (2011). A snapshot of writing instruction in middle and high schools. *English Journal*, *100*, 14-27.

Arum, R., & Roska, J. (2011). *Academically adrift: Limited learning on college campuses*. University of Chicago Press.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, *4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Aull, L. L. (2017). Corpus analysis of argumentative versus explanatory discourse and its implications for writing task design. *The Journal of Writing Analytics*, *1*(1), 1-47.

Bahr, P. R., Fagioli, L. P., Hetts, J., Hayward, C., Willett, T., Lamoree, D., Newell, M. A., Sorey, K., & Baker, R. B. (2019). Improving placement accuracy in California's community colleges using multiple measures of high school achievement. *Community College Review*, *47*(2), 178-211.

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed-response and other performance assessment*s. Educational Testing Service. https://www.ets.org/Media/About_ETS/pdf/8561_ConstructedResponse_guidelines.pdf

Beigman Klebanov, B., Burstein, J., Harackiewicz, J. M., Priniski, S. J., & Mulholland, M. (2017). Reflective writing about the utility value of science as a tool for increasing STEM motivation and retention–Can AI help scale up? *International Journal of Artificial Intelligence in Education*, *27*(4), 791-818.

Beigman Klebanov, B., & Flor, M. (2013). Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp.1148-1158).

Beigman Klebanov, B., Gyawali, B., & Song. Y. (2017). Detecting good arguments in a non-topic-specificway: An oxymoron? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)* (pp. 244-249).

Klebanov, B. B., Leong, C. W., & Flor, M. (2015). Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP* (pp. 11-20).

Beigman Klebanov, B., Leong, C. W., Gutierrez, E. D., Shutova, E., & Flor, M. (2016). *Semantic classifications for detection of verb metaphor*s. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 101-106).

Bennett, R. E., & Ward W. C. (2009). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Routledge.

Bolchover, D. (2012). *Competing across borders: How cultural and communication barriers affect business*. The Economist Intelligence Unit. https://www.ibs-b.hu/documents/278/Economist_Intelligence_Unit__Competing_across_borders_2012.pdf

Brown, D. W., & Aull, L. L. (2017). Elaborated specificity versus emphatic generality: A corpus-based comparison of higher- and lower-scoring Advanced Placement exams in English. *Research in the Teaching of English*, *51*(4), 394-417.

Burstein, J., Beigman Klebanov, B., Elliot, N., & Molloy, H. (2016). A left turn: Automated feedback & activity generation for student writers. In *Proceedings of Language Teaching, Learning and Technology* (pp. 6-13).

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online service. *AI Magazine*, *25*(3), 27-36.

Burstein, J., Elliot, N., Klebanov, B. B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing Mentor™: Writing progress using self-regulated writing support. *The Journal of Writing Analytics*, 2, 285-313. https://wac.colostate.edu/docs/jwa/vol2/bursteinetal.pdf

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998, August). Automated scoring using a hybrid feature identification technique. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (Vol. 1, pp. 206-210). Association for Computational Linguistics.

Burstein, J., McCaffrey, D., Beigman Klebanov, B., & Ling, G. (2017). *Exploring relationships between writing and broader outcomes with automated writing evaluation*. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Burstein, J., McCaffrey, D., Beigman Klebanov, B., Ling, G., & Holtzman, S. (2019). *Exploring writing analytics and postsecondary success indicators*. In *Companion Proceedings 9th International Conference on Learning Analytics & Knowledge (LAK19)* (pp. 213-214).

Burstein, J., McCaffrey, D., Elliot, N., Beigman Klebanov, B., Molloy, H., Houghton, P., & Mladineo, Z. (2020). Exploring writing achievement and genre in postsecondary writing. In *Companion Proceedings in the 10th International Conference on Learning Analytics & Knowledge (LAK20)* (pp.53-55).

Burstein, J., Tetreault, J., Chodorow, M., & Blanchard, D. (2013). Automated evaluation of discourse coherence quality in essay writing. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of Automated Essay Evaluation* (pp. 289-302). Routledge.

Cahill, A., & Evanini, K. (2020). Natural language processing for writing and speaking. In D. Yan, A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring* (pp. 69-92). Chapman and Hall/CRC.

Cassidy, L., Yee, K., Schmidt, R., Vasquez, S., Means, B., & Krumm, A. (2016). C*lassroom trials: A study of instruction with writing software*. SRI International. https://www.sri.com/wp-content/uploads/pdf/classroom-trials_jan_29_2016_1.pdf

Chen, C. E., & Cheng, W. E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, *12*, 94-112.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*(1), 100-108.

Cotos, E., Huffman, S., & Link, S. (2015). Furthering and applying move/step constructs: Technology-driven marshalling of Swalesian genre theory for EAP pedagogy. *Journal of English for Academic Purposes*, *19*, 52-72.

Council of Writing Program Administrators. (2014). *WPA Outcomes Statement for First-Year Composition (v3.0).* https://wpacouncil.org/aws/CWPA/asset_manager/get_file/350909?ver=3890

Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project. (2011). *Framework for success in postsecondary writing*. https://wpacouncil.org/aws/CWPA/asset_manager/get_file/350201?ver=7548

Cramer, E., Little, M. E., & McHatton, P. A. (2018). Equity, equality, and standardization: Expanding the conversations. *Education and Urban Society*, *50*(5), 483-501. https://doi.org/10.1177%2F0013124517713249

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7-24. https://doi.org/10.1016/j.asw.2012.10.002

Deane, P., Wilson, J., Zhang, M., Li, C., Li, C., van Rijn, P. Guo, H., Roth, A., Winchester, E., & Richter, T., (2021). The sensitivity of a scenario-based assessment of written argumentation to school differences in curriculum and instruction. *International Journal of Artificial Intelligence*, *31*, 57-98.

Department of Education. (2021). *Education in a pandemic: The disparate impacts of COVID-19 on America's students*. Office for Civil Rights. https://www2.ed.gov/about/offices/list/ocr/docs/20210608-impacts-of-covid19.pdf

Federal Interagency Working Group on Improving Measurement of Sexual Orientation and Gender Identity in Federal Surveys. (2016). *Current measures of sexual orientation and gender identity in federal surveys*. National Center for Education Statistics. https://nces.ed.gov/FCSM/pdf/buda5.pdf

Foltz, P. (2014). Improving student writing through automated formative assessment: Practices and results. In *Proceedings of the 2014 International Association for Educational Assessment Annual Conference* (pp. 1-10).

Foltz, P. W., & Rosenstein, M. (2017). Data mining large scale formative writing. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics and educational data mining* (pp. 199-201). Society for Learning Analytics Research. https://www.solaresearch.org/wp-content/uploads/2017/05/hla17.pdf

Gere, A. R., Curzan, A., Hammond, J. W., Hughes, S., Li, R., Moos, A., Smith, K., Van Zanen, K., Wheeler, K. L., & Zanders, C. J. (2021). Communal justicing: Writing assessment, disciplinary infrastructure, and the case for critical language awareness. *College Composition and Communication*, *72*(3), 384-412.

Godden, D. M., & Walton, D. (2007). Advances in the theory of argumentation schemes and critical questions. *Informal Logic*, *27*(3), 267-292.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers*, *36*(2), 193-202.

Graham, S., Bruch, J., Fitzgerald, L., Furgeson, J., Greene, K., Kim, J. S., Lyskawa, J., Olson, C. B., & Wulsin, C. S. (2016). *Teaching secondary students to write effectively*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/508_WWCPG_SecondaryWriting_122719.pdf

Graham, S., & Harris, K. (2014) Conducting high quality writing intervention research: Twelve recommendations. *Journal of Writing Research*, *6*(2), 89-123. https://doi.org/10.17239/jowr-2014.06.02.1

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, *8*(6), 1-44.

Gubala, C., Larson, K., & Melonçon, L. (2020). Do writing errors bother professionals? An analysis of the most bothersome errors and how the writer's ethos is affected. *Journal of Business and Technical Communication*, *34*(3), 250-286.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement*, *55*(2), 194-216.

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*(5), 745-765.

Hardy, J. A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-Level Student Papers (MICUSP). *Corpora*, *8*(2), 183-207.

Hart Research Associates. (2015). *Falling short? College learning and career success: Selected findings from online surveys of employers and college students conducted on behalf of the Association of American Colleges & Universities*. https://www.aacu.org/sites/default/files/files/LEAP/2015employerstudentsurvey.pdf

Hazelton, L., Nastal, J., Elliot, N., Burstein, J., & McCaffrey, D. F. (2021). Formative automated writing evaluation: A standpoint theory of action. *Journal of Response to Writing*, *7*(1) 37-91.

Hoang, G. T., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly*, *13*, 359-376.

Kaufer, D., Geisler, C., Vlachos, P., & Ishizaki, S. (2006). Mining textual knowledge for writing education and research: The DocuScope project. In L. van Waes, M. Leijten, & C. Neuwirth (Eds.), *Writing and digital media* (Vol. 17, pp. 115-129). Brill.

Kelly-Riley, D. (2015). Toward a validational framework using student course papers from common undergraduate curricular requirements as viable outcomes evidence. *Assessing Writing*, *23*, 60-74.

Ling, G., Elliot, N., Burstein, J. C., McCaffrey, D. F., MacArthur, C. A., & Holtzman, S. (2021). Writing motivation: A validation study of self-judgment and performance. *Assessing Writing*, *48*, 1-15 .

Liu, O. L., Frankel, L, & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current states and directions for next-generation assessment* (RR-14-10). ETS.

Madnani, N., Burstein, J., Sabatini, J., Biggers, K., & Andreyev, S. (2016). Language Muse: Automated Linguistic Activity Generation for English Language Learners. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics—System Demonstrations* (pp. 79–84).

Matsumura, L., Wang, E., Correnti, R., & Litman, D. (2020, July 28). *What do teachers want to see in automated writing evaluation systems*? The RAND Blog. https://www.rand.org/blog/2020/07/what-do-teachers-want-to-see-in-automated-writing-evaluation.html

Melzer, D. (2014). *Assignments across the curriculum: A national study of college writing*. Utah State University Press.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mislevy, R. J, & Elliot. N. (2020). Ethics, psychometrics, and writing assessment: A conceptual model. In J. Duffy & L. Agnew (Eds.), *After Plato: Rhetoric, ethics, and the teaching of writing* (pp. 143-162). Utah State University Press.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.). (2008). *Assessment, equity, and opportunity to learn*. Cambridge University Press.

National Center for Education Statistics. (2012). *Writing 2011: National Assessment of Educational Progress at grades 8 and 12*. U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

Palmquist, M. (2019). Directions in writing analytics: Some suggestions. *The Journal of Writing Analytics*, *3*, 1-12.

Perin, D., & Lauterbach, M. (2018). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, *28*(1), 56-78.

Rhodes, T. L. (2021). Accreditation for learning: The multi-state collaborative to advance opportunities for quality learning for all students. In D. Kelly-Riley & N. Elliot (Eds.), *Improving outcomes: Disciplinary writing, local assessment, and the aim of fairness* (pp. 225-237). Modern Language Association.

Ringler, H., Beigman Klebanov, B., Kaufer, D. (2018). Placing writing tasks in local and global contexts: The case of argumentative writing. *The Journal of Writing Analytics*, *18*, 34-77.

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*(4), 1010-1025.

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

Shum, S. B., & Crick, R. D. (2016). Learning analytics for 21st century competencies. *Journal of Learning Analytics*, *3*(2), 6-21. https://doi.org/10.18608/jla.2016.32.2

Shum, S. B., Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016). Critical perspectives on writing analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16*) (Vol. 25-29, pp. 481-483). Association for Computing Machinery. https://doi.org/10.1145/2883851.2883854

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In B. E. Clauser & M. B. Bunch (Eds.), *The history of educational measurement* (pp. 111-135). Routledge.

Sireci, S. G., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admissions Test scores. *Educational and Psychological Measurement*, *6*6, 305-317.

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International conference on computational linguistics: Technical papers* (pp. 950-961).

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Languae Science Press.

Stevenson, M., & Phakiti, A. (2013). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, *19*, 51-65.

Wetzel, D., Brown, D., Werner, N., Ishizaki, S., & Kaufer, D. (2021). Computer-assisted rhetorical analysis: Instructional design and formative assessment using DocuScope. *The Journal of Writing Analytics*, *5*, 292-323. https://doi.org/10.37514/JWA-J.2021.5.1.09

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94-109.

Woods, B., Adamson, D., Miel, S., & Mayfield, E. (2017, August). Formative essay feedback using predictive scoring models. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2071-2080).