Exploring Relationships Between Writing & Broader Outcomes With Automated Writing Evaluation

Jill Burstein Dan McCaffrey Beata Beigman Klebanov Guangming Ling
{jburstein, dmccaffrey, bbeigmanklebanov, gling}@ets.org

Educational Testing Service
Princeton, NJ 08541

Abstract

No significant body of research examines writing achievement and the specific skills and knowledge in the writing domain for postsecondary (college) students in the U.S., even though many at-risk students lack the prerequisite writing skills required to persist in their education. This paper addresses this gap through a novel *exploratory* study examining how automated writing evaluation (AWE) can inform our understanding of the relationship between postsecondary writing skill and broader indicators of college success. The exploratory study presented in this paper was conducted using test-taker essays from a standardized writing assessment of postsecondary student learning outcomes. Findings showed that for the essays, AWE features were found to be predictors of *broader outcomes* measures: college success indicators and learning outcomes measures. Study findings expose AWE's potential to support educational analytics -- i.e., relationships between writing skill and broader outcomes –moving AWE beyond writing assessment and instructional use cases.

1 Introduction

Writing is a challenge, especially for at-risk students who may lack the prerequisite writing skills required to persist in U.S. 4-year postsecondary (college) institutions (NCES, 2012). Educators teaching postsecondary courses that require writing could benefit from a better understanding of writing achievement and its role in postsecondary success (college completion). U.S K-12 research examines writing achievement and the specific skills and knowledge in the writing domain (Berninger, Nagy & Beers, 2011; Olinghouse, Graham, & Gillespie, 2015). No parallel significant body of research exists for postsecondary students. There has been research related to essay writing on standardized tests and college success indicators for exams, such as the College Board Advanced Placement[1] (Bridgeman & Lewis, 1994). However, only the final overall essay score is evaluated. In this work, we try to *drill deeper* into essays to explore if specific features in the writing of college students is related to measures of broader outcomes.

Automated writing evaluation (AWE) systems typically support the measurement of pertinent writing skills for automated scoring of large-volume, high-stakes assessments (Attali & Burstein, 2006; Shermis et al, 2015) and online instruction (Burstein et al, 2004; Foltz et al, 2013; Roscoe et al, 2014). AWE has been used primarily for on-demand essay writing on standardized assessments. However, the real-time, dynamic nature of NLP-based AWE affords the ability to explore linguistic features and skill relationships across a range of writing genres in postsecondary education, such as, on-demand essay writing tasks, argumentative essays from the social sciences, and lab reports in STEM courses (Burstein et al, 2016). Such relationships can provide educational analytics that could be informative for various stakeholders, including students, instructors, parents, administrators and policy-makers.

This paper discusses an *exploratory* secondary data analysis, using AWE to examine interactions between writing and broader outcomes measures of student success. An evaluation was conducted using test-taker essays from a standardized writing assessment of postsecondary student learning outcomes. Findings suggested that AWE features from the essays were found to be predictors of broader outcomes measures: *college success indicators* and *learning outcomes measures*. Recent

[1] https://apstudent.collegeboard.org/home

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 101–108
Copenhagen, Denmark, September 8, 2017.

work has shown similar results, examining relationships between AWE and read ing skills (Allen et al, 2016) versus broader outcomes measures

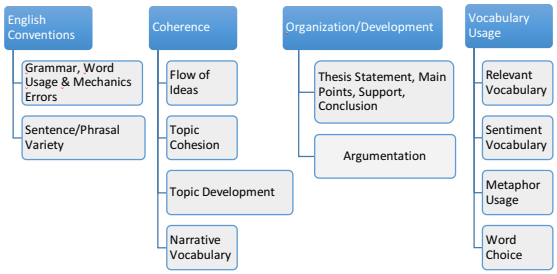

Figure 1. Construct representation of the AWE features extracted from pilot study essays.

(discussed here).

The work presented here broadens the lens -- exposing AWE's potential to inform our understanding of the relationship between writing and critical educational outcomes above and beyond prevalent use cases for assessment and instruction of writing itself.

2 The Study

An *exploratory* secondary data analysis was conducted to examine relationships between responses to a 45-minute, timed standardized *writing assessment* of postsecondary student learning. The writing assessment contains two components: an *on-demand essay task* requiring students to compose an essay in response to a prompt wherein they must adopt or defend a position or a claim presented in the prompt; and 15 *selected-response* (SR) (multiple choice) items related to one reading passage. The SR portion measures writing domain knowledge skills, such as English conventions, vocabulary choice, evaluating evidence, analyzing arguments, understanding the language of argumentation, evaluating organization, distinguishing between valid and invalid arguments, and evaluating tone. The *writing assessment* is one of three component skills assessments from an outcomes assessment suite. A second *critical thinking* component test is also used for this study. It is also a 45-minute, timed assessment, composed of 27 or 29 selected-response items depending on the test form (i.e., version of a test). The pilot study includes 5 forms (versions) for the *critical thinking* test. The five forms were developed under the same test specification and their scores were linked to each other and can be used interchangeably (Liu, et al., 2016).

In this study, we examine relationships between AWE features found in essay responses of 4-year postsecondary students who took the writing assessment, and indicators of college success.

2.1 Data

To evaluate the psychometric properties of the assessment and to gather evidence on the reliability and validity of the test prior to its release, the authors' organization had previously conducted an extensive pilot test of the assessment at more than 33 colleges and universities. Analyses used all data collected from 929 students (37% first-year, 29% sophomores, 16% junior, and 18% seniors) enrolled at the institutions; students had completed one of two pilot forms of the *writing assessment*. Of the 929 students, 514 also had scores from the pilot *critical thinking assessment*.

In addition to the *writing assessment* essay text, the pilot test data includes human ratings for the essay responses, and selected-response items scores. We also had access to students' college GPA and some external measures such as, the *critical thinking assessment scores*, SAT[2] or ACT[3] scores, high school grade point average (GPA). Although these variables were missing for subsamples of students.

2.2 Methods

Several hundred AWE features were generated for the essay writing data. These features were drawn from a large portfolio of features used for analysis of student writing (including features from a commercial essay scoring engine). As this was an initial exploratory analysis, one of the authors selected an initial, manageable set of 61 construct-relevant features related to subconstructs, including English writing conventions (e.g., errors in grammar and mechanics), coherence (e.g., flow of ideas), organization and development, vocabulary, and topicality. See Figure 1 (above). The author hypothesized that this 61-feature subset would have strong predictive potential based on the subconstruct that each feature was intended

[2] https://collegereadiness.collegeboard.org/sat

[3] http://www.act.org/

Feature Name	Subconstruct Class	NLP-Based Feature / Resource Description
argumentation	argumentation	Detection of sentences containing argumentation (Beigman Klebanov et al, 2017)
dis_coh1	coherence	Aggregate discourse coherence quality measure (Somasundaran et al, 2014)
gen_max_lsa	coherence	Latent semantic analysis values computed for long-distance sentence pairs (Somasundaran et al, 2014)
dis_coh2, dis_coh3, dis_coh4	coherence	Three measures related to topic distribution in a text (Beigman Klebanov et al, 2013; Burstein et al, 2016)
fphajnp	collocation	Noun phrase collocations identified using a rank-ratio based collocation detection algorithm trained on the Google Web1T n-gram corpus (Futagi et al, 2008)
logdta	discourse	Aggregate value based on length of essay-based discourse element (Attali & Burstein, 2006) derived from a discourse structure detection method that identifies essay-based discourse elements (e.g., thesis statement) (Burstein et al, 2003)
grammaticality	English conventions	Aggregate value generated for relative grammaticality (Heilman et al, 2014)
logg	English conventions	Aggregate value from a set of 9 automatically-detected *grammar* error feature types (Attali & Burstein, 2006)
nsqm	English conventions	Aggregate value from a set of 12 automatically-detected *mechanics* error feature types (Attali & Burstein, 2006)
nsqu	English conventions	Aggregate value from a set of 10 automatically-detected *word usage* error feature types (Attali & Burstein, 2006)
statives	narrativity	Count measures using a manually-compiled list of *stative* verbs (i.e., express states vs. action, e.g., *feel*).
PR1, PR2	personal reflection	Aggregate scores generated related to use of personal reflection language (Beigman Klebanov et al, 2017)
complexnp	phrasal complexity	Noun phrases identified with a hyphenated adjective or a prepositional phrase modifier using regular expressions defined on constituency parses.
svf	sentence variety	Aggregate value generated based on sentence-type factors (Burstein et al, 2013)
topicdev	topic development	Detection of main topics and related words (Beigman Klebanov et al , 2013; Burstein et al, 2016)
nwf_median	vocabulary sophistication	Aggregate measure generated related to word frequency (Attali & Burstein (2006)
wordln_2	vocabulary sophistication	Aggregate measure generated related to average word length for all words in a text (Attali & Burstein, 2006)
variants1, variants2	vocabulary usage	Detection of morphologically complex inflectional (variants1) and derivational (variants2) word forms using an algorithm that first over-generates variants using rules and then filters using co-occurrence statistics computed over Gigaword. (Madnani et al, 2016)
metaphor	vocabulary usage	Detection of metaphor (Beigman Klebanov et al (2015); Beigman Klebanov et al (2016)
sentiment	vocabulary usage	Count measures based on VADER[4] sentiment lexicon entries.
vocab_richness	vocabulary usage	Aggregate feature composed of a number of text-based vocabulary-related measures (e.g., morphological complexity, relatedness of words in a text). This work is not yet published.
colprep	vocabulary usage	Aggregate measure related to collocation and preposition use (described in Burstein et al, 2013).

Table 1: The 26 Features, Subconstructs & Methods

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

to address, and its alignment with the writing assessment construct Before modeling the interactions between the 61 AWE features and other measures, an analysis was conducted to identify features that were functionally related or strongly correlated to remove redundant features. This analysis identified 35 features that were monotonic functions of other features (e.g., one feature equaled the log of a second features), very highly linearly correlated, or have very small variance. Among features that were functionally related or highly correlated, the feature most highly correlated with human ratings of the essay were retained. The outcome of this analysis was the set of 26 features listed in Table 1 (below). Only the 26 features in this subset were used for the analysis reported here.

The analysis consisted of linear regression analyses with the AWE features as the independent (or predictor) variables and scores on the critical thinking assessment, SAT or ACT, *writing assessment* selected-response (SR) items, and college GPA as the dependent variables. Separate regression analyses were conducted for each dependent variable. For example, there was a model predicting GPA as a function of *argumentation*, another model predicting GPA as function of *dis_coh1*, another model predicting GPA as a function of *gen_max_lsa*, and so on for each of the features. This modeling process was repeated for each of the dependent variables. The goal of the analysis was to determine how strongly each feature was related to each outcome. However, since better writers will probably get better scores on other tests too, we wanted to know if the features contained unique information for predicting the dependent variables, above and beyond how well the essay was written. That is, we wanted to know if two students who appear to be comparable writers based on human scores can be further differentiated by the additional properties of their writing as captured by AWE. Therefore, for each dependent variable, a series of regression models were fit that predicted the dependent variable not only as a function of each of the feature values, but also included the length of the essay and the average of the human ratings on a 6-point scale (where 1 indicates the lowest proficiency and 6, the highest). The regression models included these two additional predictors because both are related to the quality of the essay. Essay length is generally a good predictor of human ratings of essays and related to many AWE features (Chodorow & Burstein, 2004). By including these two additional predictors in the model, we were better able to isolate the relationship between the features and the dependent variable distinct from quality of the essay.

3 Results

Tables 2 to 8 (below) present the results of the regression analyses for each of the 6 outcomes. For presentation purposes, the table for each dependent variable includes only those features where the coefficient for that feature was significantly greater than zero with a p-value less than 0.05. Across all the dependent variables, 25 of the 26 variables appear in the table for one or more dependent variables. Only one feature, *metaphor*, did not emerge from the analyses. Given that 26 features were tested for each dependent variable, there is a considerable chance that p-values below 0.05 were sometimes due to chance and did not indicate a statistically significant relationship. Controlling for multiple comparisons would be required to reduce the probability of spurious p-values of less than 0.05. P-values were used to reduce the size of the tables and focus on features with the strongest evidence of a relationship with each dependent variable.

Each row contains a standardized coefficient from a model that included 3 features: (1) the AWE feature, (2) the square root of the number of words (length), and (3) the raw average of 2-3 human ratings per essay. In addition to the coefficient for the AWE feature and its standard error, the table includes the overall R-squared (R^2) for the three independent variables (AWE feature, length, and average human rating) and the *part* of the R-squared attributable to the AWE features (Inc. R^2). The R^2 measures the variance explained by the predictor.

All features in the tables explain some amount of variance showing promise of relationships between AWE features and college success and learning outcomes. Results show that for all outcomes, a breadth of features emerge, covering the *English conventions, coherence* or *argumentation*, and v*ocabulary* subconstructs. Features

shown in *italics* in Tables 2-8 indicate relatively stronger predictors (i.e., greater explained variance), using Inc. R^2 of 0.05 as a "cutoff". *Vocabulary sophistication* ("wordln_2") and *vocabulary usage* ("vocab_richness") were the stronger predictors of the *critical thinking assessment* scores, the SAT/ACT Composite Score and SAT Verbal Score. *Vocabulary usage* ("sentiment") was a stronger predictor in ACT Science.

4 Discussion and Future Work

This *exploratory*, secondary data analysis illustrates that 1) writing can provide meaningful information about student knowledge related to broader outcomes (college success indicators and learning outcomes measures) and 2) AWE has greater potential for educational analytics above and beyond current prevalent uses for writing assessment and instruction. Vocabulary features were the most consistent and strongest predictors. This is not surprising since most of the college success predictors used in this study involved intensive reading, and vocabulary knowledge is shown to be related to reading comprehension (Qian & Schedl, 2004; Quinn et al, 2015). The detailed analyses illustrated in Tables 2 – 8 do show statistically significant relationships between the full set of writing skill feature measures and broader outcomes. The big picture is that this line of research could inform instructional curriculum, assessment development, and educational policy vis-à-vis the improvement of college student success factors.

The *limitations* of this project are the small size of the data set since students were missing some of the dependent variables, and the examination of writing data from a single writing genre – i.e., on-demand essay writing. However, these will be addressed in next steps, in Fall 2017-Spring 2018. The authors will conduct a larger study with seven 4-year postsecondary partner institutions. A larger sample of student writing will be collected from approximately 2,000 students from the sites. Student writing data collected will include not only on-demand essay writing, but students will each also provide multiple authentic writing assignments from their courses. Both writing and disciplinary courses will be included in the study. Student success factor data, such as, SAT and ACT scores, college GPA, course grades, and course completion, will also be collected. We will administer the same *writing assessment* and *critical thinking assessment* to our outcomes measures. Using the new data, we will apply knowledge from this study to continue to evaluate how AWE can provide analytics related to broader outcomes measures. Further, this larger data set will span different genres which will afford the opportunity to 1) replicate this exploratory study on the same writing assessment as a baseline, and 2) apply current and enhanced analyses to authentic writing data collected from college students.

AWE has traditionally been used for writing assessment (automated essay scoring), and writing instruction (automated feedback about writing). The work presented in this paper explores new territory, and brings awareness to the potential impact of NLP in a bigger educational space – i.e., to support understanding of relationships between writing and broader outcomes of student success.

Variable	Coeffcient	Std. Error	R^2	Inc. R^2
logg	0.10	0.04	0.22	0.01
nsqu	0.17	0.04	0.24	0.02
nsqm	0.11	0.04	0.22	0.01
svf	0.27	0.06	0.25	0.03
nwf_median	0.18	0.04	0.24	0.03
wordln_2	*0.25*	*0.04*	*0.27*	*0.06*
PR1	-0.08	0.04	0.22	0.01
fphajnp	0.08	0.04	0.22	0.01
complexnp	0.12	0.04	0.23	0.01
variants1	0.23	0.04	0.26	0.04
vocab_richness	*0.27*	*0.05*	*0.26*	*0.05*
dis_coh1	0.40	0.13	0.23	0.01
sentiment	0.15	0.04	0.23	0.02

Table 2: *Critical Thinking* Composite Score; Baseline R^2 with human rating and length = 0.21

Variable	Coefficient	Std. Error	R^2	Inc. R^2
nsqu	0.12	0.03	0.23	0.01
nsqm	0.21	0.03	0.25	0.04
svf	0.11	0.04	0.22	0.01
wordln_2	0.19	0.03	0.24	0.03
grammaticality	0.12	0.03	0.22	0.01
colprep	0.08	0.03	0.22	0.01
dis_coh3	-0.10	0.03	0.22	0.01
dis_coh4	-0.11	0.05	0.22	0.00
fphajnp	0.11	0.03	0.22	0.01
complexnp	0.08	0.03	0.22	0.01
variants2	0.13	0.03	0.22	0.01
vocab_richness	0.13	0.03	0.22	0.01
dis_coh1	0.23	0.09	0.22	0.01
sentiment	0.06	0.03	0.22	0.00
statives	-0.13	0.03	0.23	0.02

Table 3: *Writing Assessment* Selected Response Score; Baseline R^2 with human rating and length = 0.21

Variable	Coefficient	Std. Error	R^2	Inc. R^2
logg	0.09	0.04	0.17	0.01
nsqu	0.10	0.04	0.17	0.01
nsqm	0.17	0.04	0.18	0.03
svf	0.25	0.05	0.19	0.03
nwf_median	0.14	0.04	0.18	0.02
wordln_2	*0.25*	*0.04*	*0.21*	*0.06*
grammaticality	0.08	0.04	0.16	0.01
colprep	0.10	0.04	0.17	0.01
PR1	-0.12	0.04	0.17	0.01
PR2	-0.12	0.04	0.17	0.01
fphajnp	0.13	0.04	0.18	0.02
complexnp	0.12	0.04	0.17	0.01
variants2	0.20	0.04	0.19	0.03
gen_max_lsa5	-0.12	0.06	0.16	0.01
vocab_richness	*0.31*	*0.04*	*0.22*	*0.06*
dis_coh1	0.26	0.12	0.16	0.01
sentiment	0.17	0.04	0.19	0.03

Table 4: SAT/ACT Composite Score (ACT rescaled to the SAT Scale); Baseline R^2 with human rating and length = 0.16

Variable	Coefficient	Std. Error	R^2	Inc. R^2
logg	0.11	0.04	0.18	0.01
nsqu	0.14	0.04	0.18	0.02
nsqm	0.15	0.04	0.18	0.02
svf	0.29	0.06	0.21	0.04
nwf_median	0.15	0.04	0.19	0.02
wordln_2	*0.29*	*0.04*	*0.24*	*0.07*
grammaticality	0.11	0.05	0.17	0.01
colprep	0.12	0.05	0.18	0.01
argumentation	0.13	0.05	0.18	0.01
PR1	-0.15	0.04	0.19	0.02
PR2	-0.12	0.05	0.18	0.01
fphajnp	0.11	0.05	0.17	0.01
complexnp	0.12	0.05	0.18	0.01
variants1	0.13	0.05	0.18	0.01
variants2	0.22	0.05	0.20	0.04
gen_max_lsa5	-0.13	0.06	0.17	0.01
vocab_richness	*0.33*	*0.05*	*0.23*	*0.07*
dis_coh1	0.28	0.13	0.17	0.01
sentiment	0.12	0.04	0.18	0.01

Table 5. SAT Verbal Score; Baseline R^2 with human rating and length = 0.16

Variable	Coefficient	Std. Error	R^2	Inc. R^2
nsqm	0.22	0.05	0.14	0.04
svf	0.19	0.06	0.12	0.02
nwf_median	0.14	0.05	0.12	0.02
wordln_2	0.20	0.05	0.14	0.03
colprep	0.10	0.05	0.11	0.01
PR1	-0.12	0.05	0.12	0.01
PR2	-0.13	0.05	0.11	0.01
fphajnp	0.10	0.05	0.11	0.01
complexnp	0.11	0.05	0.11	0.01
variants2	0.15	0.05	0.12	0.02
gen_max_lsa	-0.16	0.07	0.11	0.01
vocab_richness	0.24	0.05	0.14	0.04
sentiment	0.18	0.04	0.13	0.03

Table 6. SAT Math Score; Baseline R^2 with human rating and length = 0.10

ACT English				
Variable	**Coefficient**	**Std. Error**	**R^2**	**Inc. R^2**
nsqu	0.11	0.05	0.16	0.01
nsqm	0.15	0.05	0.17	0.02
logdta	-0.19	0.06	0.18	0.03
svf	0.17	0.07	0.17	0.02
wordln_2	0.16	0.06	0.18	0.02
dis_coh2	0.21	0.11	0.16	0.01
argumentation	0.16	0.07	0.17	0.01
variants1	0.13	0.05	0.17	0.02
vocab_richness	0.16	0.06	0.17	0.02
sentiment	0.24	0.07	0.19	0.03
ACT Math				
Variable	**Coefficient**	**Std. Error**	**R^2**	**Inc. R^2**
svf	0.18	0.07	0.12	0.02
wordln_2	0.15	0.06	0.13	0.02
complexnp	0.16	0.06	0.13	0.02
variants2	0.15	0.06	0.12	0.02
variants1	0.15	0.06	0.12	0.02
vocab_richness	0.21	0.07	0.13	0.03
dis_coh1	0.38	0.17	0.12	0.02
sentiment	0.19	0.06	0.14	0.03
ACT Reading				
Variable	**Coefficient**	**Std. Error**	**R^2**	**Inc. R^2**
logg	0.11	0.05	0.14	0.01
svf	0.17	0.07	0.15	0.02
wordln_2	0.17	0.06	0.16	0.03
PR1	-0.11	0.05	0.15	0.01
variants1	0.16	0.06	0.15	0.02
vocab_richness	0.23	0.07	0.16	0.03
sentiment	0.20	0.06	0.17	0.04
statives	-0.14	0.05	0.15	0.02
ACT Science				
Variable	**Coefficient**	**Std. Error**	**R^2**	**Inc. R^2**
logdta	-0.14	0.07	0.09	0.01
svf	0.22	0.08	0.10	0.03
wordln_2	0.14	0.06	0.09	0.02
fphajnp	0.15	0.06	0.10	0.02
complexnp	0.16	0.06	0.10	0.02
variants1	0.17	0.06	0.10	0.02
vocab_richness	0.26	0.07	0.12	0.04
sentiment	*0.23*	*0.06*	*0.12*	*0.05*

Table 7. ACT Subject Test Scores; Baseline R^2 with human rating and length: ACT English = 0.15; ACT Math = 0.11; ACT Reading = 0.13; ACT Science = 0.08

Variable	**Coefficient**	**Std. Error**	**R^2**	**Inc. R^2**
nsqu	0.09	0.05	0.05	0.01
nsqm	0.16	0.05	0.07	0.02
wordln_2	0.13	0.05	0.06	0.02
grammaticality	0.13	0.05	0.06	0.01
argumentation	0.13	0.06	0.05	0.01
topicdev	-0.10	0.05	0.05	0.01
vocab_richness	0.12	0.05	0.05	0.01

Table 8. Cumulative GPA; Baseline R^2 with human rating and length = 0.04

Acknowledgements

Research presented in this paper was supported by the Institute of Education Science, U.S. Department of Education, Award Number R305A160115. Many thanks to Binod Gyawali, Michael Flor, and Diane Napolitano for support with this work.

References

Allen, L.K., Dascalu,M., McNamara, D.S., , Crossley, S.A, Trausan-Matu, S. (2016). In Proceedings of EDULEARN16 Conference 4th-6th July 2016, Barcelona, Spain, 5269-5279. Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 1-31.

Beigman Klebanov, B., Burstein, J., Harackiewicz, J., Prinski, S., Mullholland, M. (2017) Reflective writing as a tool for increasing STEM motivation and retention -- can AI help scale it up? *Special Issue for the International Journal of Artificial Intelligence in Education: MARWIDE: Multidisciplinary Approaches to*

Reading and Writing Integrated with Disciplinary Education (Eds. D. McNamara, S. Muresan, R. J. Passonneau, & D. Perin).

Beigman Klebanov, B., Leong, C., Gutierrez, D., Shutova, E., Flor, M. (2016). *Semantic classifications for detection of verb metaphors.* Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 101-106.

Berninger, V. W., Nagy, W., & Beers, S. (2011). Child writers' construction and reconstruction of single sentences and construction of multi-sentence texts: contributions of syntax and transcription to translation. *Reading and Writing, 24,* 151-182. doi 10.1007/s11145-010-9262-y.

Bridgeman, B. and Lewis, C. (1994). The Relationship of Essay and Multiple-Choice Scores with Grades in College Courses, *Journal of Educational Measurement,* 31(1): 37-50.

Burstein, J., Beigman Klebanov, B., Elliot, N., & Molloy, H. (2016). A Left Turn: Automated Feedback & Activity Generation for Student Writers. To appear in the *Proceedings of the 3rd Language Teaching, Language & Technology Workshop*, co-located with Interspeech, San Francisco, CA.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater® Automated Essay Scoring System. In Shermis, M.D., & Burstein, J. (Eds.), Handbook of Automated Essay Scoring: Current Applications and Future Directions. New York: Routledge.

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

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing, 18*(1), 32–39.

Chodorow, M. and Burstein, J. (2004). Beyond Essay Length: Evaluating e-rater's Performance on TOEFL Essays. TOEFL Research Report 73, RR-04-04, Educational Testing Service, Princeton, NJ.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K (2013). Implementation and applications of the Intelligent Essay Assessor. Handbook of Automated Essay Evaluation, M. Shermis & J. Burstein, (Eds.). Pp. 68-88. Routledge, NY. NY.

Futagi, Y., Deane, P., Chodorow, M., and Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. Computer Assisted Language Learning, 21(4).

Heilman, M., Cahill, A., Madnani,N., Lopez, M., Mulholland, M. and Tetreault, J. (2014) *Predicting Grammaticality on an Ordinal Scale* Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 174-180, Baltimore, MD.

Liu, O. L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing critical thinking in higher education: the *HEIghten*™ approach and preliminary evidence. *Assessment & Evaluation in Higher Education, 41*, 677-694.

Madnani, N., Burstein, J., Sabatini, J., Biggers, K., & Andreyev, S. (2016). *Language Muse™: Automated Linguistic Activity Generation for English Language Learners*. Proceedings of the Annual Meeting of the Association for Computational Linguistics, Berlin, Germany.

National Center for Education Statistics (2012). *The nation's report card: Writing 2011* (NCES 2012–470). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Olinghouse, N. G., Graham, S., & Gillespie, A. (2015). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology,* 107(2), 391-406.

Qian, D. and Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21(1), 28-52.

Quinn J.M., Wagner R.K., Petscher Y, Lopez D. (2015). Developmental Relations Between Vocabulary Knowledge and Reading Comprehension: A Latent Change Score Modeling Study. Child Development, 6(1):159-75.

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

Somasundaran, S., Burstein, J. and Chodorow, M. (2014) Lexical Chaining for Measuring Discourse Coherence Quality in Test-taker Essays, COLING 2014, Dublin, Ireland.