Automatic Student Writing Evaluation: Investigating the Impact of Individual Differences on Source-Based Writing

Püren Öncel
Department of Psychology, University of New Hampshire, Durham, NH, USA
po1023@wildcats.unh.edu

Lauren E. Flynn
Department of Psychology, University of New Hampshire, Durham, NH, USA
lef1008@wildcats.unh.edu

Allison N. Sonia
Department of Psychology, University of New Hampshire, Durham, NH, USA
asonia@wildcats.unh.edu

Kennis E. Barker
Department of Psychology, University of New Hampshire, Durham, NH, USA
keb1064@wildcats.unh.edu

Grace C. Lindsay
Department of Psychology, University of New Hampshire, Durham, NH, USA
gcl1006@wildcats.unh.edu

Caleb M. Mcclure
Department of Psychology, University of New Hampshire, Durham, NH, USA
ctm1021@wildcats.unh.edu

Danielle S. Mcnamara
Department of Psychology, Arizona State University, Tempe, AZ, USA
danielle.mcnamara@asu.edu

Laura K. Allen
Department of Psychology, University of New Hampshire, Durham, NH, USA
laura.allen@unh.edu

ABSTRACT

Automated Writing Evaluation systems have been developed to help students improve their writing skills through the automated delivery of both summative and formative feedback. These systems have demonstrated strong potential in a variety of educational contexts; however, they remain limited in their personalization and scope. The purpose of the current study was to begin to address this gap by examining whether individual differences could be modeled in a source-based writing context. Undergraduate students (n=106) wrote essays in response to multiple sources and then completed an assessment of their vocabulary knowledge. Natural language processing tools were used to characterize the linguistic properties of the source-based essays at four levels: descriptive, lexical, syntax, and cohesion. Finally, machine learning models were used to predict students' vocabulary scores from these linguistic features. The models accounted for approximately 29% of the variance in vocabulary scores, suggesting that the linguistic features of source-based essays are reflective of individual differences in vocabulary knowledge. Overall, this work suggests that automated text analyses can help to understand the role of individual differences in the writing process, which may ultimately help to improve personalization in computer-based learning environments.

LAK21, April 12–16, 2021, Irvine, CA, USA

ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448207

CCS CONCEPTS

• **Natural Language Processing**; • **Learning**; • **Artificial Intelligence**;

KEYWORDS

source-based writing, individual differences, vocabulary knowledge, machine-learning models

ACM Reference Format:
Püren Öncel, Lauren E. Flynn, Allison N. Sonia, Kennis E. Barker, Grace C. Lindsay, Caleb M. Mcclure, Danielle S. Mcnamara, and Laura K. Allen. 2021. Automatic Student Writing Evaluation: Investigating the Impact of Individual Differences on Source-Based Writing. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3448139.3448207

1 INTRODUCTION

Writing is a complex and multi-faceted activity that requires the coordination of multiple cognitive skills and knowledge sources that facilitate goal setting, problem solving, and strategically managing memory resources [1], [2], [3], [4]. Unfortunately, many students and adults struggle to acquire the skills needed to communicate effectively through text [5]. This skill deficit is driven by numerous factors, such as the lack of ample opportunities for students to practice writing and receive feedback, as well as the insufficient time available for teachers to provide writing instruction and formative feedback [6], [7], [8]. Importantly, the writing process is unique to each individual. Students bring their own set of strengths and weaknesses to a writing task, all of which potentially affect their writing processes and strategies. These individual differences can encompass a broad range of characteristics, from students' prior knowledge (e.g., vocabulary or domain knowledge) to their affective states (e.g., their engagement with the task). Indeed, many

theoretical models of writing proficiency attempt to account for the influence of individual differences, such as knowledge, skill, and working memory capacity (e.g. [9], [10], [11]).

Prior research has demonstrated that such individual differences can manifest in word use (see [12] for an overview). For instance, a study investigating how assessments of students' vocabulary knowledge could be informed by the lexical properties of students' independent (i.e., non source-based) writing suggested that the lexical properties of essays accounted for 44% of the variance in vocabulary knowledge [13]. Further, they identified 45 indices that were significantly related to vocabulary scores: regression analyses indicated that psycholinguistic word information and academic language were most predictive of vocabulary knowledge. Importantly, these results showed that individual differences in vocabulary knowledge could be detected through automated linguistic analyses of students' writing.

Despite these advancements, little is known about whether and how individual differences similarly manifest in the linguistic properties of source-based writing. The majority of studies that have evaluated source-based writing have focused on the overlap between the essay and the source material rather than on the linguistic features of the essays [14], [15]. Others have examined the similarities and differences between the characteristics of independent and source-based essays. For example, researchers [16] analyzed linguistic differences between source-based and independent written essays for the Test of English as a Foreign Language. Results indicated that lexical sophistication features were predictive of quality in both independent essays while source-based essays were predicted by more context-specific linguistic features, such as semantic overlap and pronoun use. Independent essays were also found to include words that were labeled as more sophisticated, less imageable, and more abstract than the source-based essays. These results highlighted a significant difference between the linguistic features used in source-based versus independent essays, pointing toward a need for algorithms that are specific to source-based writing assignments [17].

The current study furthers this work and examines the role of individual differences in the context of source-based writing. In particular, we investigate whether students' vocabulary knowledge can be modeled using multi-dimensional linguistic data extracted from their source-based essays. The long-term goal of this research is to incorporate models of individual differences into writing analytics tools to provide students with more individualized instruction and formative feedback. Below we provide a brief overview of Automated Writing Evaluation (AWE) systems and their need for greater personalization. We then describe the Writing Assessment Tool (WAT), which is currently being developed to provide students, teachers, and researchers with writing analytics. Finally, we detail the current study, and discuss the results in light of future AWE systems.

1.1 Personalization in AWE Systems

AWE systems have been developed to provide automatic scores on student writing [18]. These systems generally rely on Natural Language Processing (NLP) and machine learning techniques to assign essay scores based on the content and structure of student essays [19], [20], [21], [22]. The ultimate goal of AWE research is to improve student writing through the delivery of automated feedback and instruction. In addition, a large portion of research in this area has been conducted by educational companies and is aimed at providing automated scores on high-stakes standardized writing assessments [19], [23].

Traditionally, AWE systems have focused on the holistic scoring of students' writing – indeed, the majority of research in this area has focused on the accuracy of the summative feedback (e.g., 1-6 score) provided by the systems. While efficient, this focus on holistic scoring gives researchers, teachers, and students little to no information about the nuanced factors driving the successful or unsuccessful production of these essays. For example, we may have two students who receive the same numerical score on an essay but vary largely in aspects of their writing that potentially inform these scores. One student may struggle with vocabulary and need feedback on word choices, while the other may have poor sentence structure and need feedback on the syntactic features of their writing. Thus, systems need more information about individual student writers to provide better personalization. One way to achieve this goal is to assess students on a variety of measures before they begin to use the system to create more personalized models that go beyond holistic scores to account for individual differences in student writing. However, explicit assessments have been found to negatively impact the flow of writing. Thus, it is important that researchers and developers glean insights about students writing in other ways, such as through stealth assessments [24]. Models that assess and address these individual aspects of students' writing have stronger potential to inform more tailored feedback to these students. The purpose of the current study therefore is to examine how we can use NLP techniques to develop stealth assessments of individual differences that will ultimately be used to drive personalization in AWE systems.

Prior research shows that individual differences can have a large impact on students' writing. For instance, one important difference between skilled and less skilled writers is their level of reading comprehension skill. Reading and writing are tightly connected cognitive processes [25], [26], [27], [28]; therefore, students who more successfully comprehend texts tend to be better at generating high quality writing. Similarly, writing proficiency can be influenced by differences in vocabulary knowledge [25], [29], as students who have a larger vocabulary have more flexibility in the ways that they convey their ideas. Taking individual differences into account can help develop AWE models that provide more tailored feedback and instruction to students.

One limitation of AWE systems is that they primarily focus on independent writing that does not require the use of outside source material. This is important, because source-based writing is frequently assigned in classroom and standardized testing scenarios. Until recently, research in the context of AWE systems had not focused on this genre of writing. Instead, research on source-based writing has primarily stemmed from the psychology domain and focused on the cognitive processes involved in comprehending and evaluating texts from multiple sources [30]. This research has provided critical information about the contexts in which individuals appropriately evaluate source information and successfully integrate information. For example, recent strides have been made

to create rubrics for quantitatively evaluating students' usage of source evidence in their essays [31]. Substantially less focus has been placed on other aspects of these essays more specifically related to writing quality; thus, little is known about the processes involved in the production of high-quality source-based essays and the extent to which AWE models developed in the context of independent writing will generalize to this genre.

1.2 Writing Assessment Tool

To address these (and other) shortcomings, the Writing Assessment Tool (WAT; [32] is currently being developed.) WAT is an AWE system that will provide automated writing analytics to students, teachers, and researchers on multiple dimensions and genres of writing. WAT uses natural language processing (NLP) to analyze various linguistic properties of students' essays. WAT reports hundreds of linguistic indices that relate to the structure of the text, its general readability, rhetorical patterns, lexical choices, and cohesion using a combination of components that are commonly used in NLP tools [20], [33].

Student users of WAT will have the opportunity to practice their writing frequently and iteratively, with clear goals and rapid, formative, and individualized feedback. Additionally, teachers will have access to automated writing analytics that can help them more clearly examine the strengths and weaknesses of the students in their classes. They will be able to use WAT to assign writing tasks to students and choose to receive automated scores or score the essays themselves. Finally, researchers will have access to a web-based tool, a downloadable tool, and editable software, which will allow them to conduct computational analyses of writing.

An important component of WAT is that the tool will be able to provide automated summative and formative feedback on three types of essays: persuasive (independent) essays, summaries, and source-based (integrative) essays. Thus, students will have the opportunity to practice writing across multiple contexts, which will theoretically help them to develop more generalizable writing skills.

1.3 Current Study

Previous research has examined the role of vocabulary knowledge in persuasive essays; however, it is unclear whether these findings will generalize to other genres. The purpose of the current study is to conduct an initial set of analyses to determine the extent to which individual differences in student vocabulary knowledge manifest in the multi-dimensional linguistic properties of their source-based essays. In this study, linguistic properties of the essays were computed via WAT, which calculates indices related to the descriptive, syntactic, lexical, and cohesive properties of essays. Our goal was to examine relations between these indices and vocabulary knowledge. The overarching aim of this research is to use these models of individual differences to improve the personalization of the feedback provided by WAT.

2 METHODS

2.1 Participants

This study included 106 students from a large university campus in the Southwest United States. Participants had an average age of 22.6 years (range = 21-35) and 81% reported a grade level of college freshman or sophomore.

2.2 Data Collection Procedure

Individuals in this study participated in a 2-hour session and completed a battery of individual differences tasks. For the writing task, participants were given 40 minutes to read multiple provided sources and compose an argumentative essay in response to a prompt. Students were provided with one of two different sets of texts. The first set pertained to Green Living and contained six texts. The second set of texts related to Locavores, or people who advocate for eating locally grown food and contained seven texts. Students were asked to use the sources they had been provided to construct a central argument on Green Living or Locavores. Students were instructed to support their arguments with sources and to avoid merely summarizing the text sources.

2.3 Vocabulary Knowledge

Vocabulary knowledge was assessed using the Gates MacGinitie Vocabulary Test (4th ed.) [34].This test includes 45 simple sentences, each with an underlined vocabulary word. Students are asked to read the sentence and choose the word most closely related to the underlined word.

2.4 Automatic Text Analysis

Linguistic properties of students' essays were assessed via WAT. We selected 26 indices overall with approximately 5-7 indices our primary categories of interest: Descriptive, Lexical, Syntax, and Cohesion. The linguistic indices that were intended to measure essays at multiple dimensions that have been linked to essay quality in prior work were chosen [35]. For more thorough descriptions of these indices and their theoretical links, see [36].

We calculated five **descriptive** indices related to students' essays: number of words, number of letters per word, number of paragraphs, number of sentences, and number of words per sentence. **Lexical** indices related to psycholinguistic word information measures (i.e., meaningfulness, familiarity, and imageability), academic word use (i.e., COCA academic word and bigram frequency and range) and word frequency (i.e., SUBTLEXus word frequency) and age of acquisition (i.e., Kuperman age of acquisition). **Syntactic** indices included phrasal complexity (dependents per nominal, dependents per nominal subject, noun phrase elaboration, determiners), syntactic sophistication (diversity and frequency of verb argument constructions; VAC frequency) and clausal complexity (mean length of sentence, and mean length of clause). **Cohesion** indices represented local and global cohesion. Five measures were related to cohesion at the sentence level (Adjacent sentence word overlap, basic connectives, conjunction, Adjacent sentence semantic overlap, Semantic overlap across sentences). Global cohesion indices included semantic overlap of adjacent and all paragraphs.

2.5 Statistical Analysis

Our analyses investigated whether and how multiple levels of linguistic indices within students' essays significantly predict vocabulary knowledge. Correlations were used to assess multicollinearity among the indices (threshold set at $r > .90$). The indices that did

Table 1: R^2 and RMSE Values of Combined Model

Model	R^2	*RMSE*
Linear Regression	0.187	19.199
Linear SVM	0.259	17.302
Polynomial SVM	0.294	18.508
Random Forest	0.290	16.650
Gradient Boosting Tree	0.267	17.117

not show multicollinearity were retained. When two or more indices demonstrated multicollinearity, the index that correlated most strongly with vocabulary scores was retained in the analysis. Normality of the remaining indices was assessed with skew, kurtosis, and visual data inspections, and no indices were removed based on these inspections. Multiple machine learning models were trained to predict vocabulary scores from the text properties for each of the four linguistic dimensions. In total, five prediction algorithms were used: Linear Regression, Support Vector Machine (SVM), Linear, SVM Polynomial, Random Forest, and Gradient Boosting Tree. All models were evaluated using *10*-fold cross validation, repeated 15 times until every instance was used as the test set. The success of prediction was evaluated using R^2 and the error rate is presented with root-mean-square error (*RMSE*).

3 RESULTS

On average, students' essays contained 478 words (SD=156.32), 23 sentences (SD=9.53), and 6 paragraphs (SD=4.54). Vocabulary was measured on a scale from 0-100%, and the average vocabulary score for students was 71.92% (SD=18.89).

3.1 Prediction of Vocabulary Knowledge

We first tested the combined version of our models, which contained the descriptive, lexical, syntax, and cohesion indices. The best feature model was achieved with both Random Forest and Polynomial SVM which performed at rates above chance (Random Forest, R^2 = .29, *RMSE* = 16.65; Polynomial SVM, R^2 = .29, *RMSE* = 18.51; see Table 1). The best models explained approximately 29% of the variance in vocabulary knowledge.

To build on this model, we examined how each of the four feature subtypes (i.e. descriptive, lexical, syntax, and cohesion) successfully predicted the variance in the vocabulary knowledge. A summary of the R^2 *and RMSE* of the individual models is presented in Table 2. The Lexical model performed better than both combined and other feature subtypes models. This indicates that individual differences in students' vocabulary knowledge were best predicted by the lexical features of their source-based essays. Both Random Forest and Gradient Boosting Tree models accounted for over 30 percent of the variance in students' vocabulary scores.

3.2 Exploratory Feature Analysis

To illustrate the potential importance of features in the models above, in the following section we provide the Pearson correlations between indices from the four linguistic categories and vocabulary knowledge scores.

3.2.1 Descriptive Indices. Correlations between the vocabulary scores and descriptive indices yielded significant correlations between vocabulary knowledge and word length (i.e., number of letters per word) and a marginally significant relation with number of words per sentence (see Table 3). This analysis indicates that students with high vocabulary knowledge produced essays with longer words, which is a proxy for lexical sophistication.

3.2.2 Lexical Indices. Correlation analyses between the vocabulary scores and lexical indices yielded significant correlations between vocabulary knowledge and psycholinguistic word information (meaningfulness, familiarity, imageability). Age of acquisition and word frequency were also significantly correlated with vocabulary knowledge. This indicates that students with higher vocabulary knowledge produced essays with words that were less frequent and generally more abstract (see Table 4).

3.2.3 Syntactic Indices. Significant correlations between vocabulary knowledge and syntactic indices included noun phrase elaboration and use of determiners (see Table 5).

3.2.4 Cohesion Indices. Correlations between the vocabulary scores and cohesion indices yielded significant correlations between vocabulary knowledge and local cohesion (LSA2 all sentences and basic connectives). This analysis indicates that students with higher vocabulary knowledge produced essays with less local cohesion (see Table 6).

4 DISCUSSION

The purpose of this paper was to examine the relationship between students' vocabulary knowledge and linguistic features of their source-based essays. Machine learning algorithms were trained to predict students' vocabulary scores using descriptive, syntactic, lexical, and cohesion indices of their source-based essays. Five prediction algorithms were used to predict students' vocabulary knowledge. Results of the overall models showed that the Random Forest and Polynomial SVM models performed best, with each accounting for approximately 29% of the variance in vocabulary knowledge. Interestingly, our lexical subtype model performed better than this overall model and all other subtype models. This indicates that individual differences in vocabulary knowledge potentially allow writers to produce essays with more sophisticated words but do not enhance essay quality across all dimensions. Future research should investigate the predictive power of lexical indices on other individual differences in student writing.

Correlation analyses provided more fine-grained information about the relations between vocabulary knowledge and students' source-based essays. Aligned with the predictive models, the lexical indices subtype contained the largest number of highly correlated variables. Meaningfulness, imaginability, and age of acquisition indices were the most strongly related to vocabulary knowledge scores. This indicates that students with high vocabulary scores tend to use more lexically complex writing features than those of other indices. However, other subtype indices were also highly associated with vocabulary knowledge. For example, the descriptive, cohesion, and syntactic indices showed significant correlations between vocabulary knowledge and word length, syntactic complexity, and cohesion at local and global levels. Overall, the most

Table 2: R^2 and *RMSE* Values of Subtype Models

Model	*Descriptive*		*Lexical*		*Cohesion*		*Syntax*	
	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*
Linear Regression	0.204	17.684	0.271	16.864	0.128	18.891	0.105	18.842
Linear SVM	0.202	18.116	0.254	17.490	0.122	19.150	0.104	18.998
Polynomial SVM	0.172	18.065	0.289	17.667	0.195	20.450	0.145	18.646
Random Forest	0.089	20.292	**0.311**	16.413	0.108	19.530	0.133	18.559
Gradient Boosting Tree	0.122	18.935	**0.302**	17.145	0.124	21.072	0.151	20.482

Table 3: Correlations Between Vocabulary Scores and Descriptive Indices

Descriptive Index	*r*	*p*
Number of letters per word	0.28	<0.01
Number of words per sentence	0.18	0.06
Number of words	0.15	0.13
Number of paragraphs	-0.11	0.29

Table 4: Correlations between Vocabulary Scores and Lexical Indices

Lexical Index	*r*	*p*
Meaningfulness	-0.49	<0.01
Imageability	-0.40	<0.01
Age of acquisition	0.33	<0.01
Familiarity	-0.29	<0.01
Word frequency	-0.24	0.01
Academic Frequency	0.18	0.06
Academic Bigram Frequency	0.16	0.09
Academic Bigram Range	0.12	0.22
Academic Range	-0.01	0.91

Table 5: Correlations Between Vocabulary Scores and Syntactic Indices

Syntactic Index	*r*	*p*
Noun phrase elaboration component	0.25	0.01
Determiners	0.22	0.02
VAC frequency	-0.15	0.12
Nouns as Modifiers	0.15	0.13
Diversity and frequency component	0.14	0.16
Possessives	-0.14	0.16
Association strength	0.12	0.21

highly correlated indices were those in the lexical model, specifically meaningfulness (r = -0.49) and imageability (r = -0.40), indicating a strong negative relationship between vocabulary knowledge and the use of more abstract language.

This study provides a foundation on which writing analytics can be improved to provide more personalized assessment and feedback to students. Adopting a predictive modeling framework allows us to examine vocabulary knowledge as not only a correlate

Table 6: Correlations between Vocabulary Scores and Cohesion Indices

Cohesion Index	*r*	*p*
Basic connectives	-0.22	0.02
Semantic overlap across sentences	-0.22	0.03
Adjacent sentence semantic overlap	-0.18	0.07
Conjunctions	-0.18	0.07
Adjacent sentence word overlap	-0.17	0.08
Semantic overlap across paragraphs	-0.14	0.15

of writing quality but as one of many stealth assessment measures that may be incorporated into AWE systems to provide personalized feedback [37]. Current systems tend to only provide summative and formative feedback based on the quality of the submitted essay. Here, we move away from analyses that are performance based and instead focus on modeling features of the students that may be able to enhance the adaptivity of AWE systems. Future work should build upon this research by examining the ways in which individual differences and other contextual factors manifest in the language of student writers.

An additional strength of the current study is that it employs NLP techniques to analyze the linguistic properties of the students' essays. Although previous studies have investigated the role of individual differences in the writing process, they have largely relied on human judgments of essay quality or subjective human coding of specific essay elements. Here, we leveraged NLP tools to automatically calculate the surface- and discourse-level features of students' essays. These analyses afforded us the opportunity to investigate the role of vocabulary knowledge at a much finer grain size. Thus, rather than simply concluding that vocabulary is an important component in essay quality (according to certain essay rubrics), we provide support for claims that vocabulary knowledge is most strongly related to the production of essays that contain

specific types of language. Overall, these fine grain linguistic analyses can serve as powerful tools for writing researchers, as they can provide more thorough descriptions for the various components of the writing process.

Despite these strengths, this study is not without its limitations. With a limited sample size, our results cannot be readily generalized to overall student populations. Increasing the sample size would allow for increased variance with students' individual differences. Future studies should replicate our findings using a larger and more diverse sample from a wider range of demographic groups, ages, and language backgrounds (i.e., L2 status). Further, we used a relatively limited number of linguistic features in the current study. In future studies, the linguistic features should be extended to indices related to source reliance (e.g., overlap between the source text and individual essays) as well as other features that may be more specific to source-based writing.

Overall, this study represents an example of investigating student's essays from a new perspective. We mainly focused on understanding individual differences in source-based writing with using NLP and machine learning techniques. This research contributes to literature by showing that individual differences in students' vocabulary knowledge can be predicted by the linguistic features of their essays.

ACKNOWLEDGMENTS

This research was supported in part by IES Grants R305A180261 and R305A180144, as well as the Office of Naval Research (Grants: N00014-17-1-2300 and N00014-19-1-2424). Opinions, conclusions, or recommendations do not necessarily reflect the view of the Department of Education, IES, or the Office of Naval Research.

REFERENCES

[1] L. K. Allen and D. S. McNamara, "Five building blocks for comprehension strategy instruction," *Read. Comprehension Educ. Settings*, vol. 16, no. 125, 2017.

[2] L. S. Flower and J. Hayes, "A cognitive process theory of writing," *Coll. Compos. Commun.*, vol. 32, pp. 365–387, 1981.

[3] S. Graham, "Handbook of Educational Psychology," in *Handbook of Educational Psychology*, P. Winne and P. Alexander, Eds. Mahwah, NJ: Erlbaum, 2006, pp. 457–478.

[4] J. R. Hayes, "A new framework for understanding cognition and affect in writing," in *The science of writing: Theories, methods, individual differences, and applications*, C. M. Levy and S. Ransdell, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, 1996, pp. 1–27.

[5] National Commission on Writing, "The Neglected 'R."' College Entrance Examination Board, New York, 2003.

[6] J. Baer, S. Baldi, K. Ayotte, and P. J. Green, "The reading literacy of U.S. forth-grade students in an international context: Results for the 2001 and 2006 Progress in International Literacy Study (PIRLS)," *Natl. Cent. Educ. Stat. Inst. Educ. Sci. US Dep. Educ.*, 2007.

[7] National Assessment of Educational Progress, "The nation's report card: Writing 2007," 2007, [Online]. Available: nces.ed.gov/nationsreportcard/writing/.

[8] National Assessment of Educational Progress, "The nation's report card: Writing 2011," 2011, [Online]. Available: nces.ed.gov/nationsreportcard/writing.

[9] R. T. Kellogg, "Training writing skills: A cognitive developmental perspective," *J. Writ. Res.*, vol. 1, pp. 1–26, 2008.

[10] D. McCutchen, "Knowledge, processing, and working memory: Implication for the theory of writing," *Educ. Psychol.*, vol. 35, pp. 13–23, 2000.

[11] H. L. Swanson and V. W. Berninger, "Individual differences in children's working memory and writing skill," *J. Exp. Child Psychol.*, vol. 63, pp. 358–385, 1996.

[12] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological Aspects of Natural Language Use: Our Words, Our Selves," *Annu. Rev. Psychol.*, vol. 54, pp. 547–577, 2003.

[13] L. K. Allen and D. S. McNamara, "You Are Your Words: Modeling Students' Vocabulary Knowledge with Natural Language Processing Tools," Jun. 2015, Accessed: Apr. 24, 2020. [Online]. Available: https://eric.ed.gov/?id=ED560539.

[14] D. Blaum, T. D. Griffin, J. Wiley, and M. A. Britt, "Thinking About Global Warming: Effect of Policy-Related Documents and Prompts on Learning About Causes of Climate Change," *Discourse Process.*, vol. 54, no. 4, pp. 303–316, May 2017, doi: 10.1080/0163853X.2015.1136169.

[15] T. K. Landauer, "Automatic Essay Assessment," *Assess. Educ. Princ. Policy Pract.*, vol. 10, no. 3, pp. 295–308, Nov. 2003, doi: 10.1080/0969594032000148154.

[16] M. Kim and S. A. Crossley, "Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing," *Assess. Writ.*, vol. 37, pp. 39–56, Jul. 2018, doi: 10.1016/j.asw.2018.03.002.

[17] L. Guo, S. A. Crossley, and D. S. McNamara, "Prediction human judgements of essay quality in both integrated and independent second language writing samples: A comparison study.," *Assess. Writ.*, vol. 18, no. 3, pp. 218–238, 2013.

[18] L. K. Allen, M. E. Jacovina, and D. S. McNamara, "Computer-based writing instruction," in *Handbook of Writing Research*, C. A. MacArthur, S. Graham, and J. Fitzgerald, Eds. 2016.

[19] S. Dikli, "An overview of automated scoring of essays," *J. Technol. Learn. Assess.*, vol. 5, 2006.

[20] D. S. McNamara, S. A. Crossley, R. D. Roscoe, L. K. Allen, and J. Dai, "A hierarchical classification approach to automated essay scoring," *Assess. Writ.*, vol. 23, pp. 35–59, 2015.

[21] M. Shermis and J. Burstein, *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum, 2003.

[22] M. Warschauer and P. Ware, "Automated writing evaluation: defining the classroom research agenda," *Lang. Teach. Res.*, vol. 10, pp. 1–24, 2006.

[23] Y. Attali and J. Burstein, "Automated essay scoring with e-rater®V.2.," *J. Technol. Learn. Assess.*, vol. 4, no. 3, 2006.

[24] V. J. Shute, "Stealth assessment in computer-based games to support learning," in *Computer Games and Instruction*, S. Tobias and J. D. Fletcher, Eds. Charlotte, NC: Information Age Publishers, 2011, pp. 503–524.

[25] L. K. Allen, E. L. Snow, S. A. Crossley, G. T. Jackson, and D. S. McNamara, "Reading comprehension components and their relation to writing," *Annee Psychol.*, vol. 114, no. 4, pp. 663–691, 2014.

[26] J. Fitzgerald and T. Shanahan, "Reading and Writing Relations and Their Development," *Educ. Psychol.*, vol. 35, no. 1, pp. 39–50, 2000.

[27] T. Shanahan and R. J. Tierney, "Reading-writing connections: The relations among three perspectives. National Reading Conference Yearbook," *Natl. Read. Conf. Yearb.*, vol. 39, pp. 13–34, 1990.

[28] R. J. Tierney and T. Shanahan, "Research on the reading-writing relationship: Interactions, transactions, and outcomes," in *Handbook of reading research, Volume II*, New York: Longman, 1991, pp. 246–280.

[29] S. Graham and D. Perrin, "A meta-analysis of writing instructions for adolescent students," *J. Educ. Psychol.*, vol. 99, pp. 445–476, 2007.

[30] J. L. G. Braasch, I. Bråten, and M. T. McCrudden, Eds., *Handbook of multiple source use*. Routledge, 2018.

[31] Z. Rahimi, D. Litman, R. Correnti, E. Wang, and L. C. Matsumura, "Assessing Students' Use of Evidence and Organization in Response-to-Text Writing: Using Natural Language Processing for Rubric-Based Automated Scoring," *Int. J. Artif. Intell. Educ.*, vol. 27, no. 4, pp. 694–728, Dec. 2017, doi: 10.1007/s40593-017-0143-2.

[32] D. S. McNamara, S. A. Crossley, and R. D. Roscoe, "Natural language processing in an intelligent writing strategy tutoring system," *Behav. Res. Methods*, vol. 45, pp. 499–515, 2013.

[33] D. S. McNamara, L. K. Allen, and S. A. Crossley, "WAT: Writing Assessment Tool," in *Companion proceesings of the 9th international conference on learning analytics and knowledge*, Phoenix, AZ, USA, 2019.

[34] W. H. MacGinitie, R. K. MacGinitie, K. Maria, and L. G. Dreyer, "Gates-MacGinitie Reading Test (4th ed.)." The Riverside Publishing Company, 2000.

[35] D. S. McNamara, S. A. Crossley, and P. M. McCarthy, "Linguistic features of writing quality," *Writ. Commun.*, vol. 27, no. 1, pp. 57–86, 2010.

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

[37] L. K. Allen, A. D. Likens, and D. S. McNamara, "A multi-dimensional analysis of writing flexibility in an automated writing evaluation system," in *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, Sydney New South Wales Australia, Mar. 2018, pp. 380–388, doi: 10.1145/3170358.3170404.