Volume 7, Issue 1: 2014

Linguistic microfeatures to predict L2 writing proficiency: A case study in Automated Writing Evaluation

by Scott A. Crossley, **Kristopher Kyle**, **Laura K. Allen**, **Liang Guo**, & **Danielle S. McNamara**

This study investigates the potential for linguistic microfeatures related to length, complexity, cohesion, relevance, topic, and rhetorical style to predict L2 writing proficiency. Computational indices were calculated by two automated text analysis tools (Coh-Metrix and the Writing Assessment Tool) and used to predict human essay ratings in a corpus of 480 independent essays written for the TOEFL. A stepwise regression analysis indicated that six linguistic microfeatures explained 60% of the variance in human scores for essays in a test set, providing an exact accuracy of 55% and an adjacent accuracy of 96%. To examine the limitations of the model, a post-hoc analysis was conducted to investigate differences in the scoring outcomes produced by the model and the human raters for essays with score differences of two or greater (N = 20). Essays scored as high by the regression model and low by human raters contained more word types and perfect tense forms compared to essays scored high by humans and low by the regression model. Essays scored high by humans but low by the regression model had greater coherence, syntactic variety, syntactic accuracy, word choices, idiomaticity, vocabulary range, and spelling accuracy as compared to essays scored high by the model but low by humans. Overall, findings from this study provide important information about how linguistic microfeatures can predict L2 essay quality for TOEFL-type exams and about the strengths and weaknesses of automatic essay scoring models.

Introduction

An important area of development for second language (L2) students is learning how to share ideas with an audience through writing. Some researchers suggest that writing is a primary language skill that holds greater challenges for L2 learners than speaking, listening, or reading (Bell & Burnaby, 1984; Bialystok, 1978; Brown & Yule, 1983; Nunan, 1989; White, 1981). Writing skills are especially relevant for L2 learners involved with English for specific purposes (i.e., students primarily interested in using language in business, science, or the law) and for L2 learners who engage in standardized writing assessments used for admittance into, advancement within, and eventual graduation from academic programs.

Given the importance of writing to L2 learners, it is no surprise that investigating how the linguistic features in a text can explain L2 written proficiency has been an important area of research for the past 30 years. Traditionally, this research has focused on propositional information (i.e., the lexical, syntactic, and discoursal units found within a text Crossley, 2013; Crossley & McNamara, 2012), such as lexical diversity, word repetition, text length, and word frequency (e.g., Connor, 1990; Engber, 1995; Ferris, 1994; Frase, Faletti, Grant, & Ginther, 1999; Jarvis, 2002; Jarvis, Grant, Bikowski, & Ferris, 2003; Reid, 1986; 1990; Reppen, 1994). More recent studies have begun to assess lexical proficiency using linguistic features found in situational models (i.e. a text's temporality, spatiality, and causality, Crossley & McNamara, 2012) and rhetorical features more closely related to argument structure (Attali & Burstein, 2005; Attali, 2007). While research in this area continues to advance, a coherent understanding of the links between linguistic features in the text and how these features influence human judgments of writing proficiency is still lacking (Jarvis et al., 2003). There are many reasons for this dearth of understanding, chief among them being the sheer variety of topics, prompts, genres, and tasks that are used to portray writing proficiency. Another is the incongruence among

the types, numbers, and sophistication of the methods used to investigate writing proficiency, which renders it challenging to make comparisons between measures and studies (Crossley & McNamara, 2012).

This study focused specifically on how linguistic microfeatures produced in an independent writing task can explain L2 writing proficiency. Our focus is on one specific domain common to L2 writing research studies: a standardized independent writing assessment as found in the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT). These independent writing tasks require students to respond to a writing prompt and assignment without the use of secondary sources. Standardized independent writing assessments like those found in the TOEFL are an important element of academic writing for L2 learners and provide a reliable representation of many underlying writing abilities because they assess writing performance beyond morphological and syntactic manipulation. These assessments ask writers to provide extended written arguments that tap into textual features, discourse elements, style issues, topic development, and word choice and that are built exclusively from their experience and prior knowledge (Camp, 1993; Deane, 2013; Elliot et al., 2013). Our study builds on previous research in the area, but overcomes previous shortcomings by relying on advanced computational and machine learning techniques, which provide detailed information about the linguistic features investigated and the models of writing proficiency developed in this study. Although general explanations of the variables and models employed in AES systems have been described in previous publications (e.g., Enright & Quinlan, 2010), detailed explanations of the variables and scoring algorithms are uncommon in published studies on TOEFL writing proficiency because of the need to protect proprietary information (Attali, 2004; 2007; Attali & Burstein, 2005; 2006). In one of the most detailed of these published studies, for example, Enright and Quinlan (2010) provided an outline of the aggregated features included in e-rater (e.g., organization, development, mechanics, usage, etc.) and the relative weights for each of these aggregated features. While some of the aggregated features are fairly straightforward (e.g., lexical complexity, which is comprised of the microfeatures *average word length* and *word frequency*), others are much more opaque. For instance, the aggregated features "organization" and "development," which when combined are responsible for 61% of the e-rater score, are reported to be comprised of the "number of discourse elements" and the "length of discourse elements," respectively. However, little information is provided with regard to what qualifies as a "discourse element" or how these elements are computationally identified (though see Burstein, Marcu, and Knight, 2003). Thus, our goal in the current study is to contribute to the growing body of knowledge regarding the identification of microfeatures by providing a comprehensive, rigorous, and elaborative model of L2 writing quality using microfeatures, which can be used as a guide in writing instruction, essay scoring, and teacher training.

In addition, we examined the limitations and weaknesses of statistical models of writing proficiency (cf. Deane, 2013; Haswell & Ericsson, 2006; Herrington & Moran, 2001; Huot, 1996; Perelman, 2012; Weigle, 2013a). We did this by qualitatively and quantitatively assessing mismatches between our automated scoring model and human raters. This analysis provides a critical examination of potential weaknesses of automated models of writing proficiency that questions elements of model reliability while, at the same time, provides suggestions to improve such models. By noting both the strengths and weaknesses of an automatic essay scoring model, we hope to address concerns among scholars and practitioners about issues of model reliability (Attali & Burstein, 2006; Deane, Williams, Weng, & Trapani, 2013; Perelman, 2014; Shermis, in press) and construct validity (Condon, 2013; Crusan, 2010; Deane et al., 2013; Elliot et al., 2013, Haswell, 2006; Perelman, 2012).

L2 Writing

Writing in a second language (L2) is an important component of international education and business. Research into L2 writing has investigated a wide spectrum of variables that explain writing development and proficiency, including first language background (Connor, 1996), writing purpose, writing medium (Biesenbach-Lucas, Sigrun, & Wesenforth, 2000), cultural expectations (Matsuda, 1997), writing topic, writing audience (Jarvis et al., 2003), and the production of linguistic microfeatures (Crossley & McNamara,

2012; Ferris, 1994; Frase et al., 2000; Grant & Ginther, 2000; Jarvis, 2002; Reid, 1986, 1990, 1992; Reppen, 1994). The need to teach and assess L2 writing quickly and efficiently at a global scale increases the importance of AES systems (Weigle, 2013b).

The current study operated under the simple premise that the linguistic microfeatures in a text are strongly related to human judgments of perceived writing proficiency. The production of linguistic microfeatures in written text, especially in timed written texts where the writer does not have access to outside sources, reflects writers' exposure to a second language and the amount of experience and practice they have in understanding and communicating in that second language (Crossley, 2013; Dunkelblau, 1990; Kamel, 1989; Kubota, 1998). Also, unlike L1 writing, L2 writing can strongly vary in terms of linguistic production (e.g., syntax, morphology, and vocabulary) and is dependent on both writing ability and language ability (Weigle, 2013b). Thus, linguistic microfeatures in a text are reliable cues from which to judge L2 writing proficiency (although not the only cues). Common cues found in writing studies relate to propositional features, such as the lexical, syntactic, and discourse units found in the text. Researchers have used such cues to investigate L2 writing development and L2 writing constraints using longitudinal approaches (Arnaud, 1992; Laufer, 1994), approaches that predict essay quality (Crossley & McNamara, 2012; Ferris, 1994; Engber, 1995), approaches that examine differences between L1 and L2 writers (Connor, 1984; Crossley & McNamara, 2009; Reid, 1992; Grant & Ginther, 2000), approaches that examine differences in writing topics (Carlman, 1986; Hinkel, 2002; Bonzo, 2008; Hinkel, 2009), and approaches that examine different writing tasks (Cumming et al., 2005, 2006; Guo, Crossley, & McNamara, 2013; Reid, 1990). More recently, researchers have started to investigate situational cues (Zwaan, Magliano, & Graesser, 1995) related to a text's temporality, spatiality, or causality (Crossley & McNamara, 2009; 2012). Such studies provide foundational understandings about writing proficiency, the linguistic development of L2 writers, how L2 writers differ linguistically from L1 writers, and how prompt and task influence written production.

L2 Writing Proficiency

Our main interest in this paper is the examination of L2 writing proficiency. The most common approach to assessing writing proficiency is to assess relationships between linguistic microfeatures in an essay and the scores attributed to that essay by an expert human rater. In general, researchers have focused on lexical, syntactic, and cohesion microfeatures and how such features can predict essay scores.

Studies that have examined lexical features have found that higher rated L2 essays contain more words (Carlson, Bridgeman, Camp, & Wanderers, 1985; Ferris, 1994; Frase et al., 1999; Reid, 1986; 1990), use words with more letters or syllables (Frase et al., 1999; Grant & Ginther, 2000; Reid, 1986, 1990; Reppen, 1994), and demonstrate greater lexical diversity (Engber, 1995, Grant & Ginther, 2000; Jarvis, 2002; Reppen, 1994). Syntactically, L2 essays that are rated as higher quality include more subordination (Grant & Ginther, 2000) and instances of passive voice (Ferris, 1994; Grant & Ginther, 2000). From a cohesion perspective, researchers have investigated explicit connections and referential links within text. The findings from these studies do not demonstrate the same level of agreement as those that investigate lexical and syntactic features of text. For instance, some past studies have shown that more advanced L2 writers produce a greater number of connectives (Jin, 2001) and pronouns (Reid, 1992), while more recent studies demonstrate that higher rated essays contain fewer conditional connectives (e.g., *if-then*), fewer positive logical connectives, (e.g., *and*, *also*, *then*), less content word overlap, less given information, and less temporal cohesion (e.g., aspect repetition; Crossley & McNamara, 2012; Guo et al., 2013). In general, the findings from these studies indicate that linguistic variables related to lexical sophistication, syntactic complexity, and, to some degree, cohesion can be used to distinguish high proficiency L2 essays from low proficiency L2 essays.

Automatic Essay Scoring (AES)

Once it is established that linguistic microfeatures in a text can be used to separate high and low quality essays, it becomes possible to consider using such features to automatically score essays. Any computerized approach to analyzing texts falls under the field of natural language processing (NLP). NLP investigations of writing focus on how computers can be used to understand and analyze L2 written texts for the purpose of studying L2 writing development and proficiency. Prior to the development of NLP tools, such research required manually coding texts for linguistic microfeatures of interest, which is prone to errors, time consuming, and cost prohibitive (Higgins, Xi, Zechner, & Williamson, 2011). However, advances in computational linguistics have led to new techniques that allow researchers to automatically extract linguistic information from texts (Brill & Mooney, 1997; Dikli, 2006). These extraction techniques have led to the development of computer systems that can automatically provide assessments of the content, structure, and quality of written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Burstein, & Leacock, 2006). Such systems are known as Automatic Essays Scoring (AES) systems.

AES systems can assist teachers in scoring essays in low-stakes classroom assessments and can also offer students greater opportunities for writing practice and feedback (Dikli, 2006; Page, 2003). Additionally, AES systems can benefit large-scale testing services by providing automated and reliable ratings for high-stakes writing assessments, such as the Graduate Record Exam (GRE) or the TOEFL (Dikli, 2006). In both situations, AES systems reduce the demands and complications often associated with human writing assessment, such as time, cost, and reliability (Bereiter, 2003; Burstein, 2003; Myers, 2003; Page, 2003). Of course, AES systems are not without their detractors. A position statement written by the Conference on College Composition and Communications, the largest writing conference in North America, categorically opposes the use of AES systems in writing assessment (2004). Other researchers voice concerns that AES systems cannot assess the entire construct of writing because they fail to address issues of argumentation, purpose, audience, and rhetorical effectiveness, which are hallmarks of quality writing attended to by human raters (Condon, 2013; Deane, 2013; Haswell, 2006; Haswell & Ericsson, 2006; Herrington & Moran, 2001; Huot, 1996; Perelman, 2012). More importantly, AES systems are generally only successful at scoring limited writing genres such as the independent writing genre found in the TOEFL and less successful at assessing other genres such as authentic performance tasks and portfolio based-writing, which are considered more credible and valid forms of writing (as compared to large-scale commercial assessments; Condon, 2013; Elbow & Belanoff, 1986; Wardle & Roozen, 2013).

A few examples of AES systems that rely on NLP to assess writing are e-rater(R) (Burstein, 2003; Burstein, Chodorow, & Leacock, 2004), IntelliMetric (Rudner, Garcia, & Welch, 2005; 2006), Intelligent Essay Assessor (IEA; Landauer, Laham, & Foltz, 2003), and the Writing Pal (W-Pal) system (Crossley, Roscoe, & McNamara, 2013; McNamara, Crossley, & Roscoe, 2013). All of these systems provide scores for original essays through a comparison with a training set of annotated essays. Thus, the systems are based on the notion that essay quality is associated with specific and measurable groups of linguistic measures found in the text. AES methods first require human raters to code a set of essays for holistic quality as well as the presence of certain text properties, such as topic sentences, thesis statements, and evidence statements. The essays are then analyzed by the AES system along numerous linguistic dimensions related to lexical sophistication, syntactic complexity, grammatical accuracy, rhetorical features, and cohesion. This step allows the engine to extract linguistic features from the essays that can serve to discriminate higher- and lower-quality essays. In the last step, the extracted linguistic features are given weights and combined to create statistical models. These weighted statistical models can then be used to score essays along the previously selected dimensions.

Reliability and accuracy of Automated Essay Scoring systems. High agreement between AES engines and human raters has been reported in a number of studies (Attali, 2004; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; McNamara et al., 2013; Vantage Learning, 2003; Warschauer & Ware, 2006). The reported correlations for most AES systems typically range from .70 to .85 with one human rater, which is consistent with the range found between two human raters

(Warschauer & Ware, 2006). For instance, in an unpublished study (Attali, 2008) reported in Enright and Quinlan (2010) involving e-rater, two human raters reached an agreement of $r = .70$ while one human rater and e-rater reached an agreement of r = .76. Weigle (2010) reported that correlations between the scores assigned by two human raters for 772 essays written by 386 ESL students in response to two prompts ranged from .64 (topic 1) to .67 (topic 2), while the correlations between the averaged human scores and the e-rater scores ranged from .76 (topic 1) to .81 (topic 2). IntelliMetric reported mean correlations between automated scores and a human score between .83 and .84, respectively (Rudner et al., 2006). Using linguistic microfeatures taken from the computational tool Coh-Metrix, which helps power the W-Pal AWE system, McNamara et al. (2013) reported a correlation of $r = .81$ between their regression model and human scores for 240 TOEFL independent essays. Unlike models reported by e-rater and IntelliMetric, McNamara et al. provided details on the linguistic microfeatures that informed their models: the number of words in the text, the average syllables per word, noun hypernymy scores, past participle verbs, and conditional connectives.

The true agreement between human raters and AES engines is typically reported in two ways: perfect agreement and perfect-adjacent agreement. Perfect-agreement reports the number of identical scores between humans and an AES system while perfect-adjacent agreement reports the number of scores that are within one point of each other. In an investigation of the e-rater system, Attali and Burstein (2006) reported perfect agreement ranging from 46% to 58%, based on the test and grade level of examinees. Attali (2008) reported that two human raters reached 56% exact and 97% adjacent agreement, while one human rater and e-rater achieved 57% exact and 98% adjacent agreement. In a large study of TOEFL essays scores (152,000 independent essays), Ramineni, Trapani, Williamson, Davey, and Bridgeman (2012) reported a 60% exact agreement and 98% adjacent agreement between two raters, while e-rater reported 59% exact and 99% adjacent agreement with one human rater. Similarly, Rudner et al. (2006) investigated the accuracy of the IntelliMetric system across two studies and reported perfect agreements from 42% to 65% and adjacent agreement from 92% to 100%.

Method

Our purpose in this study was to examine the potential for automatic indices reported by the computational tools Coh-Metrix and WAT to predict human scores of essay quality in a corpus of independent essays written for the TOEFL.

Corpus

Our selected corpus of independent essays samples was collected from two administrations of the TOEFL-iBT. The essays were composed by two groups of 240 test-takers who were stratified by quartiles for each task (N =480). The essays were written on two different prompts (one prompt per form). The essays, the final scores, and the demographic information of the test-takers were directly provided by the Educational Testing Service (ETS). The 480 test-takers included both English as a Second Language (ESL) and English as a foreign language (EFL) learners. They were from a variety of home countries and linguistic backgrounds.

Scoring Rubric

The TOEFL independent writing rubric, which describes five levels of writing performance (scored 1 through 5), was used to score the independent essays (2008 rubric here). In the rubric, linguistic sophistication at the lexical and syntactic levels is emphasized in addition to the development and the coherence of the arguments along with syntactic accuracy. An independent essay with a score of 5 is defined as being a well-organized and developed response to the given topic, displaying linguistic sophistication and containing

only minor language mistakes. In contrast, an essay with a score of 1 has serious problems in organization, idea development, or language use.

Human Judgments

Two expert raters trained by ETS scored each essay using the standardized holistic rubrics described above. The final holistic score of each essay was the average of the human rater scores if the two scores differed by fewer than two points. Otherwise, a third rater scored the essay, and the final score was the average of the two closest scores. While inter-rater reliability scores are not provided for the TOEFL-iBT scores in the public use dataset, Attali (2008) reported that weighted Kappas for similarly double scored TOEFL writing samples were .70.

Research Instrument

The primary research instruments we used in this study were Coh-Metrix (e.g., Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012; McNamara, Graesser, McCarthy, & Cai, 2014) and the Writing Assessment Tool (WAT; Crossley et al., 2013). Both Coh-Metrix and WAT represent the state of the art in computational tools and together report on hundreds of linguistic indices related to text structure, text difficulty, rhetorical patterns, and cohesion through the integration of pattern classifiers, lexicons, shallow semantic interpreters, part-of-speech taggers, syntactic parsers, and other components that have been developed in the field of computational linguistics (Jurafsky & Martin, 2008). WAT, unlike Coh-Metrix, was developed specifically to assess writing quality. As such, it includes a number of writing-specific indices related to global cohesion, contextual cohesion, n-gram accuracy, lexical sophistication, key word use, and rhetorical features. The majority of Coh-Metrix and WAT indices are normed for text length (except raw counts such as token counts, type counts, and type-token ratio). Unlike other scoring engines such as e-rater, Coh-Metrix and WAT do not calculate errors in grammar, usage, mechanics and style, which have been important predictors of essay quality in previous studies (Enright & Quinlan, 2010). In total we selected 189 indices from Coh-Metrix and WAT that all had theoretical links to writing quality. The various linguistic constructs measured along with their associated indices and theoretical links are discussed briefly in Appendix A. For more detailed descriptions, we refer readers to Graesser et al., 2004, McNamara & Graesser, 2012, and McNamara et al., 2014.

Statistical Analysis

We first divided the TOEFL essay corpus into a training and a test set following a 67/33 split (Witten, Frank, & Hall, 2011). Thus, we had a training set of 320 essays and a test set of 160 essays. To control for prompt-based effects, we conducted a MANOVA that examined if the selected linguistic variables demonstrated significant differences between the two prompts. All variables that showed prompt-based effects were removed. We then conducted Pearson correlations to assess relationships between the selected variables and the human scores using the training set only. Those variables that demonstrated significant correlations ($p < .050$) with the human scores were retained as predictors in a subsequent regression analysis. Prior to inclusion, all significant variables were checked for multicollinearity to ensure that the variables were not measuring similar constructs. Our cut-off for multicollinearity was $r => .70$. If two or more indices were highly correlated with each other, we selected the index with the highest correlation to the human raters for inclusion in the regression and removed the other, redundant variable(s).

The selected indices were next regressed against the holistic scores for the 320 essays in the training set with the essay scores as the dependent variable and the Coh-Metrix indices as the predictor variables using a stepwise method. The derived regression model was then applied to the essays in the test sets to predict the scores. The R^2 provided us with an estimate for the amount of variance in human scores that the model explains. The model from this regression analysis was then applied to essays held back in the test set to assess how well it worked on an independent set of essays (i.e., how generalizable the model was to essays it was not trained on).

Exact and adjacent matches between the model and the human raters provided us with another means of assessing the reliability of essay scoring rubrics and automated scoring algorithms. The premise behind such an analysis is that a score that is only off by one point (i.e., adjacent accuracy) is more acceptable than a score that is off by 2 or more points (Attali & Burstein, 2006; Dikli, 2006; Rudner, Garcia, & Welch, 2006; Shermis, Burstein, Higgins, & Zechner, 2010).

Analysis

Prompt-Based Analysis

To control for prompt-based writing effects, which can affect linguistic production during writing (Crossley, Weston, Sullivan, & McNamara, 2011; Hinkel, 2002; 2003), a MANOVA was conducted using the selected Coh-Metrix and WAT indices as the dependent variables and the two TOEFL prompts as the independent variables. Of the 189 selected indices, only 59 of the indices did not demonstrate prompt-based effects (defined as $p > .05$ in the MANOVA). These 59 indices were thus candidates for inclusion into our models of essay writing quality.

Correlations with Human Ratings

Correlations were conducted between the 59 variables from Coh-Metrix and WAT that did not demonstrate prompt-based effects. Of these 59 variables, 43 demonstrated significant correlations with human scores for essay quality as found in the TOEFL dataset.

Multicollinearity Analysis

We next checked for multicollinearity (defined as $r > .700$) between the 43 variables to ensure they were not measuring similar or overlapping microfeatures (i.e., we selected one independent feature for each linguistic construct). Eighteen of the 43 variables yielded strong correlations with one another. For these 18 variables, we removed the variable that demonstrated the lowest correlation with the human scores of writing quality. For instance, *Number of words* correlated strongly with *Number of types* ($r = .836$), but, because *Number of types* exhibited a stronger correlation with ratings of essay quality than *Number of words*, *Number of types* was kept for the analysis and *Number of words* was removed. After controlling for multicollinearity, we were left with 34 variables for our regression analysis. These 34 variables are presented in Table 1 based on the strength of correlation with the human judgments of essay quality.

Table 1: ***Correlations between selected indices and human ratings of essay quality***

Index	*r*	*p*	Index	*r*	*p*
Number of types	0.680	<.001	Incidence of determiners	0.182	<.001

Frequency of spoken bi-grams	-0.546	.001	Incidence of possibility modals	<-0.164	<.001
Incidence of 'and'	0.392	.001	Incidence of split infinitives	0.158	<.001
Word familiarity content words	-.0367	.001	Word imageability every word	-0.157	<.001
CELEX written frequency for content words	-0.366	<.001	Mean of location and motion ratio scores	-0.157	<.001
Incidence of agentless passives	0.358	< .001	Total number of paragraphs in essay	0.154	< .001
Incidence of perfect verb forms	0.331	< .001	Minimal edit distance (all stems mean)	0.150	< .001
Average of word hypernymy	0.329	< .001	Incidence of hedges	0.150	< .001
Word meaningfulness all words	-0.304	< .001	Incidence of emphatics	0.148	< .001
Incidence of downtoners	0.267	< .001	Stem overlap	-0.147	< .001
LSA body to conclusion	0.254	< .001	Incidence of amplifiers	0.134	< .010
Incidence of conjuncts	0.254	< .001	Incidence of positive causal connectives	-0.122	< .010
Incidence of noun phrases	-0.243	< .001	Incidence of split auxiliaries	0.106	< .050
Number of motion verbs per verb phrases	0.224	< .001	Incidence of adjectival phrases	-0.105	< .050
Word concreteness component score	-0.216	< .001	Incidence of the verb 'seem'	0.099	< .050
Subordinating conjunctions	0.198	< .001	Incidence of body paragraph n-grams	0.098	< .050
Relative clause pronoun deletion in present participles	0.195	< .001	Proportion of key words	0.091	< .050

Regression Analyses

Training set. A stepwise regression analysis using the 34 indices as the independent variables to predict the human scores yielded a significant model, F (6, 352) = 55.176, p < .001, r = .716, R^2 = .512, for the training set. Six Coh-Metrix and WAT indices were included as significant predictors of the essay scores. The six indices were: *Number of types*, *Word imageability every word*, *Proportion of key words*, *Incidence of 'and'*, *LSA body to conclusion*, and *Incidence of perfect verb forms*.

The model demonstrated that the six indices together explained 51% of the variance in the evaluation of the 320 independent essays in the training set (see Table 2 for additional information). t-test information for the six indices together with the amount of variance explained are presented in Table 3.

Table 2: ***Stepwise regression analysis for indices predicting the independent essay scores: Training set***

Entry	Index added	*r*	r^2	*B*	B	*S.E.*
Entry 1	Number of types	0.654	0.428	0.017	0.608	0.001
Entry 2	Word imageability every word	0.676	0.456	-0.017	-0.162	0.004
Entry 3	Proportion of key words	0.696	0.485	4.339	0.146	1.303
Entry 4	Incidence of 'and'	0.703	0.495	0.036	0.116	0.013
Entry 5	LSA body to conclusion	0.710	0.504	0.173	0.114	0.064
Entry 6	Incidence of perfect verb forms	0.716	0.512	0.010	0.099	0.004

Note: *B* = unstandardized β; B = standardized; S.E. = standard error. Estimated constant term is 5.414.

Table 3: ***t-value, p-values, and variance explained for the six indices in the regression analysis: Training set***

Index	*t*	*p*	r^2
Number of types	12.325	p<.001	0.428
Word imageability every word	-4.026	p<.001	0.028
Proportion of key words	3.330	p<.001	0.028
Incidence of 'and'	2.693	p<.010	0.010
LSA body to conclusion	2.712	p<.010	0.009
Incidence of perfect verb forms	2.342	p<.050	0.008

Test set. We used the model reported for the training set to predict the human scores in the test set. To determine the predictive power of the six variables retained in the regression model, we computed an estimated score for each integrated essay in the independent test set using the B weights and the constant from the training set regression analysis. This computation gave us a score estimate for the essays in the test set. A Pearson's correlation was then conducted between the estimated score and the actual score assigned on each of the integrated essays in the test set. This correlation with its R^2 was then calculated to determine the predictive accuracy of the training set regression model on the independent data set.

The regression model, when applied to the test set, reported $r = .773$, $R^2 = .598$. The results from the test set model demonstrated that the combination of the six predictors accounted for 60% of the variance in assigned scores of the 160 essays in the test set, providing increased confidence for the generalizability of our model.

Exact and Adjacent Matches

We used the scores derived from the regression model to assess the exact and adjacent accuracy of the regression scores when compared to the human-assigned scores. For this analysis, we rounded up the essay scores to the closest integer (i.e., a score of 4.5 was rounded up to a 5). Our baseline comparison for this model was against a default score of 3 for each essay. A default score of 3 would provide an exact accuracy of 37% and an adjacent accuracy of 78%. The regression model produced exact matches between the predicted essay scores and the human scores for 263 of the 480 essays (55% exact accuracy). The model produced exact or adjacent matches for 460 of the 480 essays (96% exact/adjacent accuracy). The measure of agreement between the actual score and the predicted score produced a weighted Cohen's Kappa for the adjacent matches of .463, demonstrating a moderate agreement. A confusion matrix for the results is presented in Table 4.

Table 4: ***Confusion matrix for the total set of essays showing actual and predicted essay scores***

Actual Essay Score	Predicted Essay Scores				
	1	**2**	**3**	**4**	**5**
1	1	2	0	0	0
2	0	15	25	3	0
3	0	9	142	24	2
4	0	1	66	85	3
5	0	0	14	68	20

Post-hoc Analysis

Our model miscalculated human scores by a factor greater than one in four percent of the data (i.e., the predicted scores were beyond adjacent scores; n = 20). These 20 essays provide an opportunity to analyze model components that may influence reliability and text components that may influence human ratings that are not captured by our automated indices. The purpose of the following analysis is to explore the extent to which the misscored essays can provide valuable information concerning model reliability and human scoring. Our corpus consisted of those essays in which the human raters assigned a score of 5 and the model assigned a score of 3 (n = 14) and in which the human raters assigned a score of 4 and the model assigned a score of 2 (n = 1). We label these essays high/low. We also examined essays in which the human raters assigned a score of 3 and the model assigned a score of 5 (n = 2) and in which the human raters assigned a score of 2 and the model assigned a score of 4 (n = 3). We label these essays low/high.

Our post-hoc analysis was both qualitative and quantitative in nature. The 20 essays were scored by expert raters on linguistic features found in the TOEFL scoring rubric. These analytic scores were used to statistically assess the elements of the text that human raters found important and to examine differences between the high/low and low/high essays. The 20 essays were also assessed using the computational indices taken from the regression analysis. These linguistic features were used to examine differences between the high/low and low/high essays.

Analytic TOEFL scoring rubric. A coding scheme was developed to assess the major linguistic features found in the TOEFL scoring rubric (see Appendix B). These features included topic, task, development, coherence, and language use (i.e., syntactic variety, syntactic structure/accuracy, word choice, idiomaticity, vocabulary range, and spelling).

Human judgments. Two expert raters with over 5 years of teaching English to non-native speakers both abroad and in the United States were trained on the rubric using a training set of 25 TOEFL independent essays. The raters reached agreement on each analytical item after the first training session ($r > .70$)The raters then independently scored each of the essays in the post-hoc analysis corpus. If the raters disagreed by two or more points, the raters adjudicated the scores. After adjudication, all items except syntactic variety reported acceptable inter-rater reliability (see Table 5 for results).

Table 5: ***Inter-rater reliability statistics for the analytic features in the essay scoring rubric***

Feature	Cronbach's alpha	Pearson's *r*	Weighted Kappa
Topic	0.867	0.805	0.705
Task	0.937	0.884	0.860
Development	0.891	0.816	0.613
Coherence	0.881	0.792	0.732
Syntactic variety	0.701	0.549	0.519
Syntactic structure	0.921	0.864	0.845
Word choice	0.853	0.772	0.680

Idiomaticity	0.879	0.786	0.712
Vocabulary range	0.778	0.637	0.595
Spelling	0.895	0.835	0.810

>**Correlations between analytic features and human scores.** To assess links between the analytic scores and the holistic scores, we conducted correlations between the mean scores for the two raters on each analytic feature and the holistic scores assigned to the essays (see Table 6). The correlations demonstrated that 7 out of the 10 analytic features demonstrated significant correlations with the holistic scores. The strongest correlations were reported for spelling, appropriate word choice, syntactic structure/accuracy, and idiomaticity. Although Coh-Metrix and WAT measure lexical and syntactic structures, they do not measure the accuracy of the structures produced.

Table 6: ***Correlations between analytic features and holistic scores***

Feature	*r*	*p*
Topic	0.187	> .050
Task	0.061	> .050
Development	0.430	> .050
Coherence	0.472	< .050
Syntactic variety	0.620	< .010
Syntactic structure/accuracy	0.787	< .001
Word choice	0.819	< .001
Idiomaticity	0.672	< .001
Vocabulary range	0.595	< .010
Spelling	0.932	< .001

MANOVA high/low and low/high essays (analytic features). A MANOVA was conducted using the analytic indices as the dependent variables and the high/low and low/high categorizations as the independent variables. Seven out of the ten analytic features demonstrated significant differences between the categorizations and 8 of 10 of the analytic features showed a medium or larger effect size (Cohen, 1988; see Table 7). The seven features that demonstrated significant differences were the same that reported significant correlations between analytic features and holistic scores. The results indicate that essays scored high/low were rated as having greater coherence, greater syntactic variety, greater syntactic structure/accuracy, better word choices, better idiomatic language, greater vocabulary range, and better spelling than low/high essays. As in the correlation analysis, the strongest effect sizes were reported for spelling, word choice, and syntactic structure.

Table 7: ***MANOVA results for predicting high/low and low/high classifications using analytic essay features***

Analytic features	High/Low	Low/High	*F*	*p*	*hp2*	*Cohen's d*
Topic	5.433 (0.904)	5.100 (1.084)	0.465	> .050	0.025	0.334
Task	4.667 (1.319)	4.800 (1.441)	0.037	> .050	0.002	-0.096
Development	3.900 (0.870)	3.100 (0.548)	3.661	> .050	0.169	1.100
Coherence	4.133 (0.694)	3.000 (1.323)	6.313	< .050	0.260	1.073
Syntactic variety	4.400 (0.507)	3.600 (0.652)	8.151	< .050	0.312	1.370
Syntactic structure/accuracy	4.500 (0.627)	2.600 (0.652)	33.844	< .001	0.653	2.971
Word choice	4.467 (0.550)	2.800 (0.274)	41.36	< .001	0.697	3.837
Idiomaticity	4.200 (0.592)	2.800 (0.671)	19.746	< .001	0.523	2.213
Vocabulary range	4.467 (0.667)	3.300 (0.837)	10.171	< .010	0.361	1.542
Spelling	4.933 (0.594)	2.200 (1.323)	96.363	< .001	0.843	2.724

MANOVA high/low and low/high essays (computational indices). A MANOVA was conducted using the computational indices from the regression as the dependent variables and the high/low and low/high categorizations as the independent variables. Two out of the six indices demonstrated significant differences between the categorizations and four of the six indices demonstrated medium or larger effect sizes (Cohen, 1988; see Table 8). The two features that demonstrated significant differences were the number of types and incidence of perfect verbs. The results indicate that essays scored high/low contained fewer tokens and fewer perfect verb forms than essays scored low/high.

Table 8: ***MANOVA results for predicting high/low and low/high classifications using computational indices***

Computational indices	High/Low	Low/High	*F*	*p*	*hp2*	*Cohen's d*
Number of types	131.800 (14.447)	167.400 (32.067)	11.477	< .010	0.389	-1.431
Word imageability every word	317.721 (6.239)	318.801 (10.964)	0.077	> .050	0.004	-0.121
Proportion of key words	0.104 (0.032)	0.102 (0.017)	0.028	> .050	0.002	0.078
Incidence of 'and'	3.600 (2.586)	5.400 (2.191)	1.939	> .050	0.097	-0.751
LSA body to conclusion	0.952 (0.770)	1.484 (1.242)	1.320	> .050	0.068	-0.514
Incidence of perfect verb forms	8.841 (6.847)	18.552 (7.730)	8.141	< .050	0.311	-1.329

Discussion

Automated models of human essay scoring can provide strong evidence for how microfeatures of language found in the text can predict essay quality. The current study demonstrates that a regression model using six linguistic microfeatures related to breadth of lexical production, lexical sophistication, key words use, local and global cohesion, and tense can explain 60% of the variance in the human scores for the TOEFL essays in our test set. The same microfeatures can be used to predict human essay scores 55% of the time and provide adjacent matches 96% of the time. These six microfeatures, their calculations, and their weights in a regression analysis provide a straightforward, comprehensive, rigorous, and elaborative model of TOEFL independent writing quality that has applications in classroom teaching and assessment, teacher training, standardized testing situations, and industrial development.

A post-hoc analysis of our findings demonstrated that the essays scored high by the regression model and low by human raters contained a greater number of word types and perfect tense forms compared to essays scored high by human raters and low by the regression analysis. On the other hand, the analytic feature analysis demonstrated that essays scored high by humans but low by the regression model had greater coherence, syntactic variety, syntactic structure/accuracy, words choice, idiomaticity, vocabulary range, and spelling accuracy as compared to essay scored high by the model but low by human raters. These findings highlight potential problems with automated models of writing quality, especially those based on Coh-Metrix and WAT, and provide examples to better understand issues of model reliability (Attali & Burstein, 2006; Deane et al., 2013; Perelman, 2014; Shermis, in press) and construct validity (Condon, 2013; Crusan, 2010; Deane et al., 2013; Elliot et al., 2013, Haswell, 2006; Perelman, 2012). Conversely, the findings also afford us the opportunity to consider where scoring mismatches in human and automated approaches occur and, thus, provide convenient examples to guide the development of natural language processing tools.

Regression Analysis

The regression model demonstrates that the strongest predictor of essay quality is the number of word types used by an L2 writer, explaining about 43% of the variance in the model (see Table 2). This index is informative for a number of reasons. First, the index relates to a writer's breadth of vocabulary knowledge, with more word types indicating a greater vocabulary. Second, the number of word types also strongly correlates with the number of words in a text (r = .836) indicating that test-takers who produce more types (and thus more words) will receive a higher essay score. The next strongest predictor is word imageability scores, which explained about 3% of the variance in the regression model. The regression model indicates that test-takers who produce less imageable words will receive a higher score than those who produce more imageable words. Thus, lexical sophistication is an important predictor of L2 writing quality. The third strongest predictor is the proportion of key words in the essay, which explains about 3% of the variance in the human scores. The regression model indicates that test-takers who use more key words specific to the prompt (i.e., words commonly used by other test-takers for the same prompt) will receive a higher score. Those who use fewer key words are presumably less on topic and will receive a lower score. The next two indices in the regression analysis are related to cohesion. The first index, the incidence of 'and,' is related to local cohesion and explains 1% of the variance in the human scores. The index demonstrates that test-takers who use a greater number of 'and's are rated as high proficiency writers, presumably because they make greater connections between words. The second index of cohesion, LSA body to conclusion score, is related to global cohesion. This index explains 1% of the variance and indicates that writers who have greater semantic overlap between their body paragraphs and their conclusion paragraph will receive higher scores. The final index is the incidence of perfect forms. This index explained 1% of the variance and demonstrates that test-takers who produce more complex verb forms will be rated as more proficient writers.

Correlational Analysis

Of secondary interest are the indices that demonstrated significant correlations with human ratings (see Table 1), but were not included in the regression analysis. These correlations generally support the findings from the regression model in that better rated essays correlated with indices of lexical sophistication (e.g., contained less frequent n-grams, less familiar words, less frequent words, less meaningful words, and less concrete words). The correlations are less clear in terms of text cohesion. Some cohesion indices show positive correlations with essay quality (e.g., conjuncts and subordinating conjunctions), while others show negative correlations (e.g., stem overlap and positive causal connectives) or indicate lower cohesion through a positive correlation (minimal edit distance). The correlations seem to indicate that conjuncts and connectives are important indicators of essay quality, but overlap, causality, and minimal edit distance are not. The correlations also seem to demonstrate that essays that are more verbal (i.e., contain more perfect forms and motion verbs) are rated higher than essays that are more nominal (i.e., contain a greater incidence of noun phrases). Two other trends are evident in the correlation analysis. The first is that more syntactically complex essays are rated higher, as evidenced by the positive correlations between essay score and indices such as incidence of agentless passives, incidence of relative clause deletion, incidence of split infinitives, and split auxiliaries. The second trend is that essays with more rhetorical features, such as downtoners, hedges, emphatics, amplifiers, and body paragraph n-grams are scored higher by human raters. These correlations, along with the indices included in the regression model, provide strong indications of the types of linguistic microfeatures that predict human ratings of essay quality.

Importantly, many of these indices relate directly to textual elements that are of concern for automated models of essay quality. These include elements such as text cohesion and coherence, text relevance, and rhetorical purposes. While indices related to text length, lexical sophistication, and syntactic complexity may not overlap with writing concerns voiced by both the writing studies and educational measurement communities (e.g., domain knowledge, cultural and background knowledge, and variation in rhetorical purposes; Condon, 2013; Deane, 2013; Haswell, 2006; Haswell & Ericsson, 2006; Herrington & Moran, 2001; Huot,

1996; Perelman, 2012), they do indicate an ability to quickly and easily produce complex text, which should free up cognitive resources that can be used to address rhetorical and conceptual concerns in the text, both of which are needed for writing mastery (Deane, 2013). They thus can be used to provide empirical evidence to help represent a more robust construct representation of writing quality (Elliot & Klobucar, 2013; Kane, 2013). However, this suggestion warrants empirical investigation.

Post-hoc Analysis

Our post-hoc analysis provides an overview of the weaknesses of the tested regression model. While the adjacent accuracy reported for our regression model is on par with previous analyses of TOEFL independent essays (Attali, 2008; Rudner et al., 2006), 4% of the essays were scored outside the adjacent range with the human ratings. The majority of these essays (n = 15) were scored lower by the model than by human raters. The main reason for this scoring discrepancy appears to be the model's reliance on assigning a stronger weight to the number of word types in the essay. While essay length can be a strong indicator of an essays organization and development (Attali & Powers, 2008), this may not be the case for all writers (Crossley, Roscoe, & McNamara, 2014). The largest effect size between the high/low and low/high essays in the post-hoc analysis was for number of word types (see Table 8), with essays scored high by humans and low by the model averaging 132 words and essays scored high by the model and low by the human averaging 167 words. This finding indicates that vocabulary breadth and essay length are not always synonymous with essay quality for human raters and that the described model assigns too much importance to these microfeatures. Similar findings, based on effect sizes, are reported for the incidence of perfect verb forms, incidence of 'and,' and LSA body to conclusion scores. However, since these indices only explained about 1% of the variance each in the regression model, their weight is less predictive than that reported for word type count.

The analyses of the human ratings for analytic features provide some indications about which linguistic elements are associated with essay quality when fewer word types are used and the text is of shorter length (see Table 7). Foremost, it appears that human raters rely on spelling accuracy to assign scores in such cases. The problem is compounded by the notion that, according to the Coh-Metrix calculations, the number of tokens may increase for each misspelled word. Thus, if a test-taker spells 'the' as both 'the' and 'hte,' that essay will be judged to have a greater number of types by the model and thus be given a higher score, because the calculation for the incidence of tokens does not consider misspelled words. Conversely, the essay will be scored lower by human raters because of the increased number of spelling errors.

The next two most important analytic features for humans raters in essays scored high/low and low/high are word choice and syntactic structure/accuracy. The essays scored high by humans and low by the regression model had better word choices as compared to those essays scored low by humans and high by the model. Thus, it is not just that words are spelled correctly, but that words are used appropriately, something that neither the Coh-Metrix nor WAT indices assess (or for that matter, any AES system of which we are aware). This is a major limitation of current AES systems. In addition, those essays scored high by humans and low by the model had fewer syntactic errors as compared to those essays scored low by humans and high by the model. Again, while Coh-Metrix calculates indices that can assess syntactic complexity, it does not report indices for syntactic accuracy (although some AES systems do assess grammatical errors).

Implications for AES systems

Considering the correlations reported in Table 6 between the analytic features and the holistic scores for the essays, the findings from this paper indicate that, at a minimum, a more successful AES system should include computational indices of spelling and syntactic accuracy (at least when L2 writing is being assessed). Automated assessments of mechanical and spelling errors are linguistic microfeatures that computers are relatively accurate at capturing, so such an implementation should not be difficult (see e-

rater as an example). Other linguistic elements such as accurate word choice and idiomaticity are more difficult to implement and point to areas in which AES systems need improvement, while features such as coherence and syntactic variety are already measured by Coh-Metrix and WAT. It should be noted that the limitations discussed here are linguistic in nature and do not address larger conceptual concerns expressed by many writing researchers regarding the inability for AES systems to assess the effectiveness of written arguments, stated purposes, rhetorical moves, and addressing the appropriate audience.

Prompt Effects

The findings also provide evidence that researchers should use caution when examining writing quality in a corpus of essays written on a number of different prompts. Prompt-based effects occur when linguistic features found in a writing prompt influence the writing patterns found in essays written on that prompt (Brown, Hilgers, & Marsella, 1991; Huot, 1990). Past studies have demonstrated prompt-based differences in text cohesion (Crossley, Varner, & McNamara, 2013), syntactic complexity (Crowhurst & Piche, 1979; Hinkel, 2002; Tedick, 1990), and in lexical sophistication (Crossley, Weston, et al., 2011; Hinkle, 2002). The present analysis controlled for prompt-based differences, but reported that of 189 potential indices, only 59 did not show prompt-based differences (i.e., 69% of the indices had to be removed from the analysis because they were influenced by the prompt). Thus, prompt should be a major concern for researchers interested in AES.

Microfeatures or Component Scores?

A final issue relates to the specificity of the linguistic indices that were used to predict essay scores and their subsequent representation of the constructs within the essays. The microfeatures calculated by Coh-Metrix and WAT are, by their nature, extremely fine-grained indices, which are intended to represent certain characteristics of learners' essays. In many ways, this is a strength of such microfeatures because researchers can use the linguistic indices to investigate specific questions about language use in various forms of texts. However, the use of these fine-grained microfeatures as predictors of essay quality could potentially lead to less stable models, which do not generalize to different prompts and tasks. It is yet to be seen whether microfeatures, aggregated features (like those used by e-rater), or a combination of both are most informative and predictive. In future studies, we plan to address this issue by developing component scores based on these linguistic microfeatures. The development of such component scores may improve the stability of AES algorithms and provide more representative features, as well as provide more informative means for formative writing feedback to students.

Conclusion

Overall, this study provides important information about how linguistic microfeatures can predict L2 essay quality and about the strengths and weaknesses of automatic essay scoring models. Unlike previous research, this study provides specific information on the linguistic microfeatures that correlate with L2 essay quality and a regression model that can be used to automatically assign scores to the TOEFL essays using these linguistic microfeatures. While the results are strong, future studies should consider similar approaches using a larger corpus of data (this study was limited to the 480 essays in the TOEFL iBT public use dataset). A larger corpus is especially needed in analyses that examine model miscalculations, because, in strong models, miscalculations are infrequent, leading to potentially small sample sizes (e.g., in our study only 20 essays with mismatched scores existed).

Overall, the results of this study advance our knowledge of how linguistic features in an L2 essay predict human judgments of quality. Follow-up analyses discuss some of the weaknesses of AES systems and provide suggestions for AES system development

including the incorporation of lexical and syntactic accuracy indices. These improvements to AES systems should provide greater overlap between human and automated ratings of essay quality. Automating essay scoring should free teachers from many elements of essay grading that are time consuming and cost prohibitive, allowing them to focus more on other aspects of essay quality that AES systems are poor at assessing, such as argumentation, style, and idea development.

Acknowledgments
This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. TOEFL (R) test material are reprinted by permission of Educational Testing Service, the copyright owner.

Biography
Scott Crossley is an Associate Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, and second language acquisition. He has published articles in second lexical acquisition, second language writing, second language reading, discourse processing, language assessment, intelligent tutoring systems, and text linguistics.

References

Arnaud, P. J. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In P. J. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 133-145). London, England: Macmillan.

Attali, Y. (2004, April). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays*. Princeton, NJ: ETS.

Attali, Y. (2008). *E-rater performance for TOEFL iBT independent essays*. Unpublished manuscript.

Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater(R) v.2.0*. Princeton, NJ: ETS.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater(R) v.2.0. *The Journal of Technology, Learning and Assessment, 4*(3), (np).

Attali, Y., & Powers, D. (2008). A developmental writing scale (ETS Research Report RR-08-19). Princeton, NJ: Educational Testing Service.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (CD-ROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bereiter, C. (2003). Foreword. In Mark D. Shermis, & Jill C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary approach* (pp. vii-ix). Mahwah, NJ: Lawrence Erlbaum Associates.

Biesenbach-Lucas, S., & Weasenforth, D. (2001). E-mail and word-processing in the ESL classroom: How the medium affects the message. *Language Learning and Technology, 5*, 35-165.

Bell, J., & Burnaby, B. (1984). *A handbook for ESL literacy*. Toronto, Canada: Ontario Institute for Studies in Education/Hodder and Stoughton.

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

Bialystok, E. (1978). A theoretical model of second language learning. *Language Learning, 28*, 69-83.

Bonzo, J. D. (2008). To assign a topic or not: Observing fluency and complexity in intermediate foreign language writing. *Foreign Language Annals, 41*(4), 722-735.

Brill, E., & Mooney, R. J. (1997). An overview of empirical natural language processing. *AI Magazine, 18*, 4-13.

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. *Written Communications, 8*, 533-556.

Brown, G., & Yule, B. (1983). *Discourse analysis*. Cambridge, England: Cambridge University Press.

Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring: A cross-disciplinary approach* (pp. 113-121). Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online service. *AI Magazine, 25*, 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 Natural Language Processing, 18*(1): 32-39.

Camp, R. (1993). Changing the model for the direct writing assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45-78). Cresskill, NJ: Hampton Press, Inc.

Carlman, N. (1986). Topic differences on writing tests: How much do they matter? *English Quarterly*, 19, 39-49.

Carlson, S., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and non-native speakers of English*. (TOEFL Research Rep. No. 19). Princeton, NJ: ETS.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conference on College Composition and Communication. (2004, February 25). CCCC position statement on teaching, learning, and assessing writing in digital environments. Retrieved from http://www.ncte.org/cccc/resources/ positions/digitalenvironments

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

Connor, U. (1984). A study of cohesion and coherence in English as a second language students' writing. *Papers in Linguistics, 17*(3), 301-316.

Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second-language writing*. Cambridge, England: Cambridge University Press.

Costerman, J., & Fayol, M. (1997). *Processing interclausal relationships: Studies in production and comprehension of text*. Hillsdale, NJ: Lawrence Erlbaum Associates. Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238.

Crismore, A., Markkanen, R., & Steffensen, M. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication, 10*, 39-71.

Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching, 46*(2), 256-271.

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 214-219). Menlo Park, CA: The AAAI Press.

Crossley, S. A., & McNamara, D. S. (2009). Computationally assessing lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 17*(2), 119-135.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *IJCEELL, 21*, 170-191.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading, 35*, 115-135.

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

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2013). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In P. M. McCarthy & G. M. Youngblood (Eds.). *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 208-213). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication, 31*(2), 184-215.

Crossley, S. A, Salsbury, T., & McNamara, D. S. (2009). Measuring second language lexical growth using hypernymic relationships. *Language Learning. 59*(2), 307-334.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*(3), 573-605.

Crossley, S. A., Varner, L. K., & McNamara, D. S. (2013). Cohesion-based prompt effects in argumentative writing. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 26th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 202-207). Menlo Park, CA: The AAAI Press.

Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*(3), 282-311.

Crowhurst, M. C., & Piche, G. L. (1979). Audience and mode of discourse effects on syntactic complexity in writing on two grade levels. *Research in the Teaching of English, 13*, 101-109.

Crusan, D. (2010). *Assessment in the second language writing classroom*. Ann Arbor, MI: University of Michigan Press.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5-43.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype tasks for the new TOEFL*. TOEFL Monograph Series, Report No. 30. Princeton, NJ: ETS.

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

Deane, P., Williams, F., Weng, V. Z., & Trapani, C. S. (2013). Automated essay scoring in innovative assessments of writing from sources. *Journal of Writing Assessment, 6*(1), 40-56.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1), 4-35.

Douglas, D. (1981). An exploratory study of bilingual reading proficiency. In S. Hudelson (Ed.). *Learning to read in different languages* (pp. 33-102). Washington, D.C.: Washington Center for Applied Linguistics.

Douglas, R. D. (2013). The lexical breadth of undergraduate novice level writing competency. *The Canadian Journal of Applied Linguistics, 16*(1), 152-170

Dufty, D. F., McNamara, D. S., Louwerse, M., Cai, Z., & Graesser, A. C. (2004). Automatic evaluation of aspects of document quality. In S. Tilley & S. Huang (Eds.), *Proceedings of the 22nd Annual International Conference on Design of Communication: The Engineering of Quality Documentation* (pp. 14-16). New York: ACM Press.

Dufty, D. F., Graesser, A. C., Lightman, E., Crossley, S. A., & McNamara, D. S. (2006). An algorithm for detecting spatial cohesion in text. Paper presented at the 16th Annual Meeting of the Society for Text and Discourse, Minneapolis, MN.

Dufty, D. F., Graesser, A. C., Louwerse, M., & McNamara, D. S. (2006). Is it just readability, or does cohesion play a role? In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1251-1256). Mahwah, NJ: Erlbaum.

Dunkelblau, H. (1990). A contrastive study of the organizational structures and stylistic elements of Chinese and English expository writing by Chinese high school students. *Dissertation Abstracts International, 51*(4), 1143A.

Elbow, P., & Belanoff, P. (1986). Using portfolios to judge writing proficiency at SUNY Stony Brook. In: P. Connolly & T. Vilardi (Eds.), *New directions in college writing programs* (pp. 95-105). New York, NY: Modern Language Association.

Elliot, N., Gere, A. R., Gibson, G., Toth, C., Whithaus, C., & Presswood, A. (2013). Uses and limitations of automated writing evaluation software. *WPA-CompPile Research Bibliographies, 23*.

Elliot, N., & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis, J. Burstein, & S. Apel (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 16-35). New York, NY: Routledge.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater(R) scoring. *Language Testing, 27*(3), 317-334.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*(2), 414-420.

Fellbaum, C. (1998). *WordNet: an electronic lexical database.* Cambridge, MA: MIT Press. Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL test of written English*. (TOEFL Research Report No. 64). Princeton, NJ: ETS.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123-145.

Guo, L. (2011). *Product and process in TOEFL iBT independent and integrated writing tasks: An investigation of construct validity*. Unpublished doctoral dissertation. Georgia State University.

Guo, L. Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Writing Assessment, 18*(3), 218-238.

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

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

Haswell, R. H. (2006). Automatons and automated scoring: Drudges, black boxes, and dei ex machina. In: P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 57-78). Logan, UT: Utah State University Press.

Haswell, R., & Ericsson, P. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*, 480-499.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language, 25*(2), 282-306.

Hinkel, E. (2002). *Second language writers' text*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hinkel, E. (2003). Adverbial markers and tone in L1 and L2 students' writing. *Journal of Pragmatics, 35*(7), 1049-1068.

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics, 41*, 667-683.

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

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

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57-84.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377-403.

Jin, W. (2001). *A quantitative study of cohesion in Chinese graduate students' writing: Variations across genres and proficiency levels*. Retrieved from ERIC database (ED452726).

Jurafsky, D., & Martin, J. H., (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics* (2nd Ed). Prentice-Hall.

Kamel, G. W. (1989). *Argumentative writing by Arab learners of English as a foreign and second language: An empirical investigation of contrastive rhetoric*. Dissertation Abstracts International, A: The Humanities and Social Sciences, 677-A.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.

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

Kubota, R. (1998). An investigation of L1-L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric. *Journal of Second Language Writing, 7*, 69-100.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review, 104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2&3), 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th*

Annual Meeting of the Cognitive Science Society (pp. 412-417). Mahwah, NJ: Erlbaum.

Landauer, T. K., Laham, D., & Foltz, P. W. (2000). The Intelligent Essay Assessor. *IEEE Intelligent systems*, September/October, 27-31.

Landauer, T. K., Laham, R. D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. Shermis & J. Bernstein (Eds.), *Automated Essay Scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25*(2), 21-33.

Longo, B. (1994). Current research in technical communication: The role of metadiscourse in persuasion. *Technical Communication, 41*, 348-352.

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics, 12*, 291-315.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313-330.

Malvern, D. D., Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, England: Palgrave Macmillan. doi: 10.1057/9780230511804

Matsuda, P. K. (1997). Contrastive rhetoric in context: A dynamic model of L2 writing. *Journal of Second Language Writing, 6*(1), 45-60.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International, 66*(12), (UMI No. 3199485)

McCarthy, P. M., & Jarvis, S. (2007). *vocd*: A theoretical and empirical evaluation. *Language Testing, 24*, 459-488. doi: 10.1177/0265532207080767

McCarthy, P.M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 381-392.

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

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods, 45*(2), 499-515.

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

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

Miller, G., Beckwith, A., Fellbaum, R., Gross, C., & Miller, K. J. (1990). Introduction to WordNet: An online lexical database. *International Journal of Lexicography, 3*, 235-244.

Myers, M. (2003). What can computers and AES contribute to a K-12 writing program? In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 3-20). Mahwah, NJ: Lawrence Erlbaum Associates.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge, England: Cambridge University Press.

Page, E. B. (2002). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43-54). Mahwah, NJ: Lawrence Erlbaum Associates.

Pearson, P. D. (1974). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relationships. *Reading Research Quarterly, 10*, 155-192.

Perfetti, C. A. (1985). *Reading ability*. Oxford, England: Oxford University Press.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In: C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), *International advances in writing research: Cultures, places, measures* (pp. 121-131). Fort Collins, Colorado: WAC Clearinghouse/Anderson, SC: Parlor Press.

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104-111 .

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater(R) scoring engine* (ETS Research Report No. RR-09-01). Princeton, NJ: ETS.

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater(R) scoring engine for the TOEFL(R) independent and integrated prompts* (ETS Research Report No. RR-12-06). Princeton, NJ: ETS.

Rashotte, C. A., & Torgesen, J. K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly, 20*, 180-188.

Rayner, K., & Pollatsek, A. (1994). *The psychology of reading*. Mahwah, NJ: Lawrence Erlbaum Associates.

Reid, J. (1986). Using the Writer's Workbench in composition teaching and testing. In C. Stansfield (Ed.), *Technology and language testing* (pp. 167-188). Alexandria, VA: TESOL.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191-210). Cambridge, England: Cambridge University Press.

Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing, 1*, 79-107.

Reppen, R. (1994). A genre-based approach to content writing instruction. *TESOL Journal, 4*(2), 32-35.

Rudner, L., & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer*. Retrieved from ERIC database (ED458290).

Rudner, L., Garcia, V., & Welch, C. (2005). *An evaluation of Intellimetric™ essay scoring system using responses to GMAT(R) AWA prompts* (GMAC Research report number RR-05-08).

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetricTM essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4).

Salsbury, T., Crossley, S. A, & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. *Second Language Research, 27*(3), 343-360. DOI: 10.1177/0267658310395851.

Shermis, M. D., & Barrera, F. D. (2002). Automated essay scoring for electronic portfolios. *Assessment Update, 14*(4), 1-2.

Shermis, M. D. (in press). The challenges of emulating human behavior in writing assessment. *Assessing Writing*.

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

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N. S. Petersen (Eds.), *International encyclopedia of education* (3rd Ed.). Oxford, England: Elsevier.

Shermis, M. D., Burstein, J., & Leacock, C. (2006). Applications of computers in assessment and analysis of writing. In C. A. McArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 403-416). New York, NY: Guilford Press.

Streeter, L., Psotka, J., Laham, D., & MacCuish, D. (2004). The credible grading machine: Essay scoring in the DOD [Department of Defense].

Tedick, D. J. (1990). ESL Writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes, 9*, 123-143.

Templin, M. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.

Vantage Learning. (2003). *How does IntelliMetric™ score essay responses?* (Report No. RB-929). Newtown, PA: Vantage Learning.

Wardle, E., & Roozen, K. (2013). Addressing the complexity of writing development: Toward an ecological model of assessment. *Assessing Writing, 17*, 106-119.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157-180.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335-353.

Weigle, S. C. (2013a). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 1*, 85-99.

Weigle, S. (2013b). English as a second language writing and automated essay evaluation. In M. D. Shermis, J. Burstein, & S. Apel (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36-54). New York, NY: Routledge.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary (2nd Version). *Behavioral Research Methods, Instruments and Computers, 20*(1), 6-11.

Witten, I. H., Frank, E., & Hall, M. A., (2011). Data mining: Practical machine learning tools and techniques (3rd Ed.). San Francisco, CA: Morgan Kaufmann.

White, R. (1981). Approaches to writing. *Guidelines, 6*, 1-11.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 386-397.

Appendix A: Language indices used in analysis

Basic Text Properties

Coh-Metrix provides a variety of indices that describe the basic properties and structure of a text, such as the number of words, the number of word types (i.e., the number of unique words produced), the number of sentences, the number of paragraphs, the average length of words, and the average length of sentences.

Cohesion indices.

Causal cohesion. Coh-Metrix calculates the level of causal cohesion in a text by measuring the ratio of causal verbs to causal particles (Graesser, et al., 2004; Dufty, McNamara, Louwerse, Cai, & Graesser, 2004). The measure of causal verbs is based on the frequency count of main clausal verbs identified through WordNet (Fellbaum, 1998; Miller et al., 1990). The causal particles are counted based on a defined set of main causal verbs, such as *because* and *as a result*. Causal cohesion can reduce text comprehension difficulties as it reveals causal relationships between simple clauses, as well as between events and actions (Pearson, 1974-1975).

Connectives. Coh-Metrix calculates the incidence score for connectives in a text as the number of occurrences per 1000 words. In addition to calculating a measure of all connectives contained within a given text, Coh-Metrix provides indices on five categories of connectives (Halliday & Hasan, 1976; Louwerse, 2001): causal (*because*, *so*), contrastive (*although*, *whereas*), additive (*moreover*, *and*), logical (*or*, *and*), and temporal (*first*, *until*). Finally, Coh-Metrix contrasts positive (*also*, *moreover*) versus negative (*however*, *but*) connectives. Coh-Metrix also calculates the incidence of conjuncts and subordinating conjunctions in a text. Connectives and conjuncts increase text cohesion by explicitly linking ideas and clauses together (Crismore, Markkanen, & Steffensen, 1993; Longo, 1994).

Logical operators. The logical operators measured in *Coh-Metrix include variants of or, and, not, and if-then combinations. The logical operators in Coh-Metrix are directly related to the density and abstractness of a given text. Additionally, they have been shown to correlate with higher working memory demands. (Costerman & Fayol, 1997).*

Lexical overlap. Coh-Metrix considers four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap, and content word overlap (McNamara, Crossley, & McCarthy, 2010). Noun overlap represents how frequently a common noun of the same form is shared between two sentences. Argument overlap measures how often two sentences share nouns with common stems. Stem overlap refers to how often a noun in one sentence shares a common stem with the other words types in a second sentence. Finally, content word overlap calculates the amount of shared content words between sentences. Research has shown that lexical overlap aids in text comprehension (Douglas, 1981; Kintsch & van Dijk, 1978; Rashotte & Torgesen, 1985).

Semantic co-referentiality. Coh-Metrix uses Latent Semantic Analysis (LSA) to measure the semantic co-referentiality of a text. LSA is a statistical representation of deeper world knowledge used to assess the level of semantic cohesion within a text. LSA utilizes singular value decomposition to condense large corpora of texts into approximately 300-500 dimensions (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). The dimensions represent the degree of semantic similarity between words, which is an important indicator of text cohesion (Landauer, McNamara, Dennis, & Kintsch, 2007).

Spatial cohesion. Coh-Metrix assesses spatial cohesion with two forms of information: location information and motion information (Dufty, Graesser, Lightman, Crossley, & McNamara, 2006; Dufty, Graesser, Louwerse, & McNamara, 2006). Both motion verbs and location nouns are identified through WordNet (Fellbaum, 1998; Miller, et al., 1990). Spatial cohesion aids in the construction of a well-structured situational model (Kintsch & van Dijk, 1978) that helps convey the meaning of the text.

Temporal cohesion. Temporal cohesion is measured by Coh-Metrix in three ways: aspect repetition, tense repetition, and the combination of aspect and tense repetition. Time is represented in text through two dimensions: tense (past, present, future) and aspect (in progress versus completed). With the use of these dimensions, Coh-Metrix calculates the consistency of tense and aspect across a passage of text. As shifts in tense and aspect occur, the Coh-Metrix repetition score decreases. Thus, a low score indicates that the representation of time in a given text may be disjointed, which could have a negative consequence on the construction of a mental representation.

Structural cohesion. Coh-Metrix computes the Minimal Edit Distance (MED) for a text sample by measuring differences in the sentential positioning of content words. A high MED value indicates that content words are located in different places within sentences across the text, suggesting lower structural cohesion.

Paragraph cohesion. WAT calculates the lexical and semantic overlap between paragraphs (initial to middle paragraphs, middle paragraphs to final paragraph, and initial paragraph to final paragraph) and between the essay prompt and the essay. The semantic similarity among the paragraphs, essay, and prompt are calculated using LSA cosine values. Lexical overlap, on the other hand, is calculated using measures of key word overlap. High lexical and semantic overlap between paragraph types is related to judgments of essay coherence (Crossley & McNamara, 2011).

Verb cohesion. WAT computes verb overlap using both LSA and WordNet. Using LSA, WAT computes the average cosine between verbs in adjacent sentences. Using WordNet, WAT computes a binary score based on whether verbs in adjacent sentences share the same synonym set. These indices are indicative of the extent to which verbs (which have salient links to actions, events, and states) are repeated across the text.

Prompt cohesion. WAT calculates the cohesion between the text and the prompt by computing the semantic similarity between the prompt and the response using both LSA and key word analyses. These indices have demonstrated positive correlations with writing proficiency (Crossley, Roscoe, & McNamara, 2011).

Prompt specific key word use. WAT reports two indices of prompt-specific key word use that can be used to measure the topic development in text. These indices compute the key words found in essays written on a specific prompt and then calculate the frequency of these key words in a specific essay. These indices have demonstrated positive correlations with writing proficiency (Crossley, Roscoe, et al., 2011).

Lexical indices.

Hypernymy. Coh-Metrix uses WordNet (Fellbaum, 1998; Miller et al., 1990) to report word hypernymy values for all content words, nouns, and verbs. A hypernymy value represents the degree of specificity of a word within a conceptual hierarchy. Thus, a low hypernymy score reflects a more abstract text. Hypernymy scores are associated with lexical knowledge and production (Crossley, Salsbury, & McNamara, 2009).

Polysemy. Coh-Metrix measures the word polysemy value for all content words, nouns, and verbs using WordNet (Fellbaum, 1998; Miller et al., 1990). Polysemy refers to the number of senses associated with a word; ambiguous words have more senses. Thus, word polysemy is indicative of the level of text ambiguity. As well, it is an indicator of lexical proficiency (Crossley, Salsbury, & McNamara, 2010).

Lexical diversity. Traditional lexical diversity (LD) indices typically measure the ratio of types (i.e., unique words occurring in the text) by tokens (i.e., all instances of words); higher numbers (from 0 to 1) indicate greater lexical diversity (Templin, 1957). As these indices are highly correlated with text length, they are not reliable across texts of varying token counts (McCarthy & Jarvis, 2007). Thus, Coh-Metrix reports on more sophisticated LD indices that control for text length, including *MTLD* (McCarthy, 2005; McCarthy & Jarvis, 2010) and *D* (Malvern, Richards, Chipere, & Durán, 2004). Measures of lexical diversity relate to the range of vocabulary a writer or speaker knows.

Word frequency. Word frequency indices measure how often particular words occur in the English language. The indices reported by Coh-Metrix are taken from the CELEX database (Baayen, Piepenbrock, & Gulikers, 1993), which consists of frequencies computed from a 17.9 million-word corpus. Word frequency is an important indicator of lexical knowledge. In addition, frequent words can be accessed and decoded more quickly than less-frequent words (Perfetti, 1985; Rayner & Pollatsek, 1994).

Word information measures. Word information indices are calculated in Coh-Metrix using the MRC Psycholinguistic Database. The indices provided include measures of concreteness, familiarity, imageability, and meaningfulness (Wilson, 1988). Higher concreteness scores typically refer to words that reference objects, materials, or persons. Word familiarity refers to words that are more readily recognized, but not necessarily more frequent (compare *eat* to *while*). Imageability refers to the ease with which words evoke mental images. Word meaningfulness scores represent the strength of association between words. The four measures are highly associated with word knowledge in L2 learners (Salsbury, Crossley, & McNamara, 2011).

Academic word list. WAT provides frequency counts from the Academic Word List, which consists of 3110 words commonly found in academic writing (Coxhead, 2000). The words contained in the AWL have been linked to writing proficiency (Douglas, 2013).

Syntactic indices.

Syntactic complexity. Coh-Metrix calculates a number of indices that measure syntactic complexity. These indices include the mean number of words before the main verb, the mean number of high-level constituents (sentences and embedded sentence constituents) per word, the average number of modifiers per noun phrase, the incidence of embedded clauses, the incidence of 'that' deletion, the incidence of passives, and the incidence of infinitives. Coh-Metrix also measures syntactic similarity by calculating the uniformity and consistency of the syntactic constructions at the clause, phrase, and word level for a given text. Higher rated essays contain

more complex syntactic structures, such as the number of words before the main verb (McNamara et al., 2010) and the number of embedded clauses (Guo et al., 2013).

Part of speech tags. Coh-Metrix reports incidence scores for all the part of speech tags reported by the Penn Tree Bank Tag Set. These include word and phrase level tags for content items such as noun types, verb types, adjective types, and adverb types and function items such as pronouns, determiners, demonstratives, modals, and prepositions (Marcus, Santorini, & Marcinkiewicz, 1993). These tags have been used to distinguish writing proficiency in previous studies (Guo, 2011; Crossley, Roscoe, et al., 2011).

N-grams indices.
N-gram accuracy. WAT assesses the n-gram accuracy of written text by comparing the normalized frequency of n-grams (bi-grams and tri-grams) shared in both a reference corpus taken from the British National Corpus (BNC) and the language sample of interest. The indices report correlations that represent the similarity between the frequency of occurrences in a representative corpus and a sample text. Higher rated essays contain n-grams that occur at similar frequencies as the representative corpus (Crossley, Cai, & McNamara, 2012).

N-gram frequency. WAT assesses the frequency of n-grams (bi-grams and tri-grams) found in a sample text. Higher proficiency writers use less frequent n-grams (Crossley et al., 2012).

N-gram proportion. WAT reports n-gram values (bi-grams and tri-grams) based on proportion scores. More proficient writers produce essays that contain proportionally fewer n-grams (Crossley et al., 2012).

Rhetorical features.
Paragraph specific n-grams. WAT reports on a variety of n-gram indices that are specific to the positioning of paragraphs in writing samples (i.e., introductory, body, and concluding paragraphs). These use of quality paragraph n-grams have demonstrated positive correlations with writing proficiency (Crossley et al., 2012; McNamara, et al., 2013).

Lexical features. WAT reports on a variety of lexical categories related to rhetorical style. These include amplifiers, private verbs, hedges, indirect pronouns, exemplification, copula verbs (be, appear, and seem), private and public verbs, and downtoners. Such features are important elements of written discourse (Biber, 1988)

Appendix B
Analytical rating form

Read each essay carefully and then assign a score on each of the points below. For the following evaluations, you will need to use a grading scale between 1 (minimum) and 6 (maximum).

We present here a description of the grade as a guide using the example of *does not meet the set criterion in any way* versus *meets the set criterion in every way*. For example, a grade of 1 would relate to not meeting the criterion in any way, and a grade of 4 would

relate to somewhat meeting the criterion. The distance between each grade (e.g., 1-2, 3-4, 4-5) should be considered equal. Thus, a grade of 5 (*meets the criterion*) is as far above a grade of 4 (*somewhat meets the criterion*) as a grade of 2 (*does not meet the criterion*) is above a grade of 1 (*does not meet the criterion in any way*).

Score	Definition
1	Does not meet the criterion in any way
2	Does not meet the criterion
3	Almost meets the criterion but not quite
4	Meets the criterion but only just
5	Meets the criterion
6	Meets the criterion in every way

Copyright © 2018 - ***The Journal of Writing Assessment*** - All Rights Reserved.