Castledown

OPEN ACCESS

Language Education & Assessment
ISSN 2209-3591
https://www.castledown.com/journals/lea/

*Language Education & Assessment, 2* (3), 110-134 (2019)
https://doi.org/10.29140/lea.v2n3.152

# The Effect of Cohesive Features in Integrated and Independent L2 Writing Quality and Text Classification

RURIK TYWONIW [a]
SCOTT CROSSLEY [b]

[a] *Georgia State University, USA*
Email: rtywoniw1@gsu.edu

[b] *Georgia State University, USA*
Email: scrossley@gsu.edu

## Abstract

Cohesion features were calculated for a corpus of 960 essays by 480 test-takers from the Test of English as a Foreign Language (TOEFL) in order to examine differences in the use of cohesion devices between integrated (source-based) writing and independent writing samples. Cohesion indices were measured using an automated textual analysis tool, the Tool for the Automatic Assessment of Cohesion (TAACO). A discriminant function analysis correctly classified essays as either integrated or independent in 92.3 per cent of cases. Integrated writing was marked by higher use of specific connectives and greater lexical overlap of content words between textual units, whereas independent writing was marked by greater lexical overlap of function words, especially pronouns. Regression analyses found that cohesive indices which distinguish tasks predict writing quality judgments more strongly in independent writing. However, the strongest predictor of human judgments was the same for both tasks: lexical overlap of function words. The findings demonstrate that text cohesion is a multidimensional construct shaped by the writing task, yet the measures of cohesion which affect human judgments of writing quality are not entirely different across tasks. These analyses allow us to better understand cohesive features in writing tasks and implications for automated writing assessment.

**Keywords:** cohesion; integrated writing assessment; L2 writing; TAACO; text classification.

## Introduction

Connectedness and organization have long been considered important aspects of second language (L2) writing development, L2 writing instruction, and assessment of L2 writing quality (de Silva, 2015; Grabe & Kaplan, 1996; Sasaki, 2000; Shaw & Weir, 2007). Key to a text's connectedness are the ideas of cohesion, the level of a writer's use of explicit lexical, syntactic and textual features to connect ideas throughout a text, and coherence, the level of connectedness evident to a reader of a text (Halliday & Hasan, 1976). Specifically, measures of cohesion have been utilized as a parameter for judgments of


***Data Availability Statement:*** *All relevant data are within this paper.*

writing quality alongside ratings of organization and language accuracy and sophistication in L2 writing assessment (de Silva, 2015; Grabe & Zhang, 2016; Sasaki, 2000). However, cohesion in writing, although necessary, does not sufficiently guarantee coherence for readers (Carrell, 1982). Research into the impact of cohesive feature use on human judgments of writing quality have found mixed results, with the potential for cohesive features to predict writing quality dependent on whether they measure local (i.e., sentence level cohesion) or global aspects of cohesion (i.e., cohesion linking larger text segments; Crossley, Kyle, & McNamara, 2016a; Crossley & McNamara, 2011a; Guo, Crossley, & McNamara, 2013).

In this study we use cohesion features to distinguish two writing assessment tasks (integrated and independent writing) and predict writing quality in the Test of English as a Foreign Language (TOEFL) internet-Based Test (iBT) writing samples. Many examinations of cohesive features in learner writing categorize features by their shape and function (i.e., surface features such as connectives and propositional reference; Hall et al., 2015; Halliday & Hassan, 1976; Roman et al., 2016). Other research emphasizes the categorization of cohesive features based on the distance between textual connections, with cohesive features being either local or global (Crossley, Kyle, & McNamara, 2016b; Dascalu et al, 2015; Horning & Kraemer, 2013; Ruegg & Sugiyama, 2013). Local cohesion refers to those features which are explicit in the text and connect nearby textual elements, including connectives and lexical overlap between adjacent sentences. Global cohesion refers to those features which are more implicit and capture lexical overlap across paragraphs or larger text segments rather than immediate proximities. Additionally, cohesive features may exist in a grey area between local and global, such as when repetitions of lexical items appear at both the local and global level or when specific lexical categories related to discourse organization (e.g., connectives, pronouns, or determiners) appear throughout a text. Crossley and his colleagues (2016b) refer to this overall text connectedness as text cohesion.

Examinations of cohesion in L2 writing often conceptualize the use of cohesive features as a facet of language proficiency and discourse awareness (Liu & Braine, 2005; Yang & Sun, 2012). Thus, writing assessments such as the TOEFL iBT provide fertile grounds to examine interactions of cohesion with discourse orientation and writing quality. Yet, when examining the quality of L2 writing in an assessment such as the TOEFL iBT, the influence of text type (i.e., integrated or independent writing samples) must be considered. Much previous research has examined differences in these text types using large scale computational analysis of register variation, finding that integrated and independent writing tasks differ along multiple dimensions of surface features of writing (Biber & Gray, 2013; Cumming et al., 2005; Kyle & Crossley, 2016). Previous studies have also shown that elements of cohesion to differ between types of texts (Plakans & Gebril, 2017), and that cohesion features are predictive of writing quality (Guo et al., 2013). However, no research to this point has investigated whether cohesive features which differ across writing tasks are also predictive of writing quality. Combining the foci of previous research, the goal of this study is to determine if cohesive features, at the local, global, and textual levels, can distinguish between the two TOEFL writing tasks and if a model can be built to determine which distinguishing features also predict writing quality. To this end, we use an automated text analysis tool, the Tool for the Automatic Analysis of Cohesion (TAACO), to measure the use of cohesive features in L2 writing samples from the TOEFL independent and integrated writing tasks. We then analyze which cohesion features can be used to distinguish the two tasks, and subsequently which of these distinguishing features are predictive of writing quality. This computational approach allows us to test the assumption that local, global, and text cohesive features are important components of both text types and text quality.

### Cohesion and Cohesive Features

Researchers have historically used multi-faceted analysis of surface linguistic features to understand register differences and differentiate text types (Biber, 1988; Grabe, 1987). Such analyses have shown that language learners show preferences for certain discourse features depending on the writing task (Lux & Grabe, 1991). Importantly, it is the co-occurrence of linguistic features, rather than specific lexical items, that differentiate written text types. In terms of cohesion features, this co-occurrence goes deeper than the correlation of distinct linguistic features, and involves features of cooccurrence, such as lexical repetition, pronominal reference, and given-new relations (Halliday & Hassan, 1976; Hoey, 1991). Many cohesion features are prime examples of surface features that are not limited to semantic categories yet form an integral part of a text's discourse orientation.

Research interested in cohesion in discourse focuses on links between both sentences and larger segments of text (i.e., local and global cohesion respectively). A text is cohesive when there are linguistic features that link ideas between sentences and aid in creating a unified texture (Halliday & Hasan, 1976). Cohesive features exist explicitly at the lexical level as well as at the relational level (i.e., between lexical items). These features help writers direct the organization of text, mark ideas as new or given, and give cues to relationships between referential objects (Halliday & Hasan, 1976). Explicit cohesive devices include isolated features, like connectives (e.g., and, however), and the connections between sections of text, usually at the lexical level. For example, Hoey (1991) specifically highlighted repetition of lexical items across text as an early-acquired method for writers to signal the connectivity of a text.

The locality of connection established by cohesive devices is also important to their function (Crossley, Kyle, & McNamara, 2016a; 2016b; Ruegg & Sugiyama, 2013). Local cohesive features are features which establish cohesion through conjunctive expressions, discrete lexical connectives (therefore, however), and lexical overlap in local environment (e.g., the same noun subject appearing in sequential sentences). Beyond local cohesion, researchers are also interested in global cohesion cues (more implicit and meaning-based) between larger text segments (Crossley, Kyle, & McNamara, 2016b; McNamara et al., 2014). These relations and cues take the form of presuppositions or words, phrases, and structures which call back to known entities from earlier in the text, such as referential pronouns, demonstratives, comparatives, as well as anaphoric substitutions and ellipses. Additionally, textual cohesion beyond discrete text segments, whether near or far, can be measured throughout an entire text using measures such as overall lexical repetition through type-token ratios (Jarvis, 2017) or givenness, which can be measured using use of determiners (Ekiert & Han, 2016) or more complex algorithms (Hempelmann et al., 2005).

Researchers are interested in how cohesion affects text quality judgements, investigating specific lexical items, general lexical and semantic overlap across text segments, and text level coreference (Crossley, Kyle, & McNamara, 2014; 2016; Crossley & McNamara, 2012; Myers et al., 2011; Reppen, 1994). However, cohesion does not necessarily engender coherence (Carrell, 1982), and simply having a preponderance of cohesive devices like connectives does not always make for clear, coherent writing.

Cohesive strategies and features are likely to vary by text type and register (Halliday & Hasan, 1976; McNamara, Louwerser, McCarthy, & Graesser, 2010). Different types of text will contain different amounts and forms of cohesion. This notion is supported in research examining various text type dyads and specific cohesive features. For instance, Graesser et al. (2007) found monologic texts to have more referential cohesion than dialogs. In comparing authentic versus simplified texts, Crossley, Allen, and McNamara (2012) found that authentic texts have more logical connectives while simplified texts had more overall cohesion features in the form of local lexical overlap and more common connectives.

Additionally, deceptive texts were reported to have higher levels of cohesion in terms of semantic overlap between sentences and amount of given information when compared to truthful texts (Duran et al., 2010). Reynolds (2001) found lexical repetition of second language writers to be dependent on writing task (and the resulting text type) with persuasive writing involving more lexical repetition than descriptive writing. These studies indicate that cohesive features vary depending on text types and that those features can be used to distinguish text types.

### Cohesion in Language Assessment

Various studies have compared cohesive features to first language (L1) writing quality and L2 writing quality with mixed results. L1 research indicates that the use of cohesive features in younger writers is related to writers' knowledge of organizational and self-regulation strategies in writing (Garcia & Fidalgo, 2008; Koutsoftas & Peterson, 2016) and measure writing development. Indeed, cohesive features have been used to assess development of writing form, discourse organization and genre awareness in L1 (Galloway & Uccelli, 2015; Reynolds & Perin, 2009). For instance, Bunch and Willet (2013) found cohesion to develop along with writing development, although they only examined the larger construct of organization. Galloway and Uccelli (2015) found use of cohesion features to develop through early grade levels, but that it was a distinct construct with low correlation with lexico-grammatical skill. Research with high school writers has shown that the uptake and use of cohesive devices is correlated with business letter writing quality (Duggleby, Tang, & Kuo-Newhouse, 2015), and high school students' comprehension of academic texts increases with the presence of cohesive features (Hall et al., 2015; Reed & Kershaw-Herrera, 2016). However, some research suggests that the use of such cohesive markers has minimal effects and is invariant between grade levels (Pinto, Tarchi, & Bigozzi, 2015). Crossley and McNamara (2010, 2011a), explicitly distinguishing local and global cohesion features, found that while the use of explicit, local cohesion features was negatively correlated with L1 writing quality, increased global cohesion was an indicator of writing quality. Crossley and his colleagues (2016) were able to identify three global cohesion features that were not text-based but rather measured based on semantic similarity which positively correlated with writing quality. In sum, what type of cohesion features appear in a text is more important for judgments of writing quality than how many explicit linking words are used. The use of local cohesive devices is just one path to building a text which may be coherent for the reader, so it should not be surprising that local cohesion devices may not be strongly related to ratings of proficiency (Guo et al., 2013; Staples & Reppen, 2015).

L2 researchers, teachers, and assessment specialists have also been interested in L2 writing and text cohesion finding that the use of cohesive devices signals writing development, discourse competence, and writing quality (Liu & Braine, 2005; Yang & Sun, 2012). Similar to studies in L1 writing research, cohesive features have been used as a measure of L2 writing quality in L2 middle grade writing (Bunch & Willet, 2013). Teachable features of cohesion are often included as targets for teaching discourse, writing, and genre (Celce-Murcia & Olshtain, 2000; McCarthy & Carter, 1994), but direct links between use of cohesion features and perceived writing quality may not be readily apparent (Watson Todd et al., 2007). Cohesion features have also been included as evidence of writing quality in language assessments, but these cohesion features are often discrete and macroscopic, with assessment instruments examining cohesion through features such as paragraph organization and use of connectives (Mullis & Mellon, 1980; Sasaki, 2000).

Cohesion figures into writing quality rubrics used in L2 assessment. For instance, cohesion is an analytic category on the IELTS writing rubrics for the integrative description task and the prompted essay task, with raters asked to address a text's organization and progression of ideas, as well as how logical, unobtrusive, and non-repetitive the cohesive devices employed are (British Council, 2018).

The only difference between how cohesion is rated between the two tasks is that the prompted essay rubric has additional attention to paragraph structure and formatting. Cohesion is not found explicitly in the TOEFL iBT rubrics for integrated and independent writing, so cohesive features may only be indirectly part of a rater's assessment. Rather, the independent writing rubric specifically asks raters to consider unity, progression, and coherence with lower rated essays showing greater redundancy and digressions and fewer connections, while the integrated rubric is generally concerned with connections made between the source and response text (Educational Testing Service, 2012).

Like L1 research, studies which have empirically examined links between writing quality and the incidence of cohesion features in L2 writing report mixed results. For instance, Duggleby, Tang, and Kuo-Newhouse (2015) found the use of certain connector words to be moderately and positively correlated with writing quality in English as a Second Language (ESL) writing. Guo et al. (2013) found that human ratings of L2 independent and integrated writing quality could be predicted using a number of linguistic features, including cohesion features. For example, semantic similarity of words-in-context between adjacent sentences as measured by Latent Semantic Analysis (Foltz, 2007) was positively correlated with human ratings of integrated writing quality. In terms of independent writing tasks, only conditional connectives, a surface-level feature, were predictive of writing quality. However, the coefficients were negative, indicating that the use of more local cohesive devices was negatively related to writing quality. Crossley, Kyle, and McNamara (2016) found numerous local cohesive devices to be either weakly correlated (e.g., certain connective words), or negatively correlated (local pronoun repetition) with L2 writing quality. Wilson, Roscoe, and Ahmed (2017), modeled L2 writing quality with Structural Equation Modeling and found that global cohesion was an integral part of L2 writing structure. In another study employing structural equation modeling, Kim and Crossley (2018) examined global overlap (between paragraphs), local overlap (between sentences), and connectives as measures of cohesion in TOEFL writing, finding that global lexical overlap across paragraphs was greater in independent writing and local lexical overlap between sentences was greater in integrated writing. No differences were reported in the use of connectives between the tasks. In terms of writing quality, they also found that while global overlap positively correlated with the scores in both tasks, local overlap only correlated positively in integrated writing, and that use of connectives was negatively correlated with the quality in independent writing.

### Current Study

The current study focuses on fine-grained indices of cohesion related to deep and surface meaning which we use to classify texts by an assessment task in the TOEFL iBT as well as measure the relationship between cohesion and human ratings of writing quality. Previous research has used features related to word frequency, collocational patterns, and lexico-grammar in L2 writing during writing assessment to explain register differences between the types of texts produced in the TOEFL iBT (Biber & Gray, 2013; Cumming et al., 2005). In addition, previous studies have analyzed the relationship between fine-grained indices of lexical sophistication and human ratings on independent and source-based writing (Kyle & Crossley, 2016), as well as lexical, syntactic, and cohesive features together on both integrated and independent writing scores (Guo et al., 2013). However, no study has looked specifically at fine-grained cohesive indices separately from other indices to differentiate learner writing by task type in the TOEFL and has compared the impact of cohesive strategy in both TOEFL iBT writing task formats on human ratings of writing quality. The current study continues the above three strands of research by applying fine-grained measurement of cohesion and discourse orientation features to classify texts by register and to connect these measurements of register-specific cohesion to writing quality.

This study analyzed 480 independent and 480 integrated expert-rated TOEFL iBT essays across two

prompts for each task. Feature counts for various cohesion indices were collected using TAACO (Crossley, Kyle, & McNamara, 2016b). Indices used for analysis include indices of semantic overlap between adjacent and distant sentences and paragraphs, use of repeated content words, use of lexical cohesive devices such as pronouns, conjunctions, subordinators and other clausal and sentential connectives, and determiners and demonstratives. The goal of this study is to understand whether measurements of cohesion in L2 writing assessment distinguishes text types and predicts scores. Specifically, these measurements were used to address two research questions:

1. Is the use of cohesive features in writing related to the type of text being written?
    a. Which cohesion features differ significantly between independent and integrated writing?
    b. Can the use of cohesive features in independent and integrated writing tasks be used to predict task type?
2. What is the relationship between measurements of cohesive features and integrated and independent writing quality as measured by human-rater judgements?

Taken together, these results will better inform second language writing researchers and writing assessment researchers regarding the strategies which L2 English writers employ to give their writing cohesion, and how cohesion can be operationalized for rating second language writing in different registers.

## Methods

### Corpus of Independent and Integrated Writing

A corpus of 960 TOEFL iBT essays from 480 test-takers was analyzed in this study. The data came from the Test of English as a Foreign Language internet Based Test (TOEFL iBT) Public Use Data Set. The TOEFL iBT includes two different writing tasks: independent writing and integrated writing. The independent prompts ask test-takers to write an essay that asserts and defends an opinion on a particular topic based on their life experience. The integrated prompt asks test-takers to read a short passage, listen to a related lecture, and synthesize the information given in the reading and the lecture. Currently, the TOEFL iBT uses both formats to assess test-taker writing proficiency.

This corpus was constructed by the Educational Testing Service (ETS) using two separate administrations of the test, each with 240 test-takers. Both administrations utilized the two writing tasks, but with different prompts for each administration. Each test taker completed both writing tasks, giving a total of 960 texts and 480 unique authors. Each combination of task and prompt was represented by 240 texts. Table 1 outlines the texts in the corpus.

Included in the corpus are the original ratings for each essay, which were rated by human raters using a holistic rubric on a scale from 1 to 5 (5 being most proficient). For both tasks, the rubric includes attention to essay coherence, clarity, and organization, all of which may be elicited by the author via cohesive features. For the integrated task, the rubric also gives attention to the selection of information integrated from the source to the essay. In the independent task, the rubric emphasizes argumentation in place of source-use constructs. The rubrics used in the data set for each task can be found on the ETS webpage (2004).

**Table 1** *Texts used in this study*

| Task type | Prompt | Number of texts |
|---|---|---|
| First administration | | |
| Integrated | Bird Migration | 240 |
| Independent | Study Subjects | 240 |
| Second administration | | |
| Integrated | Fish Farming | 240 |
| Independent | Cooperation | 240 |
| | | Total: 960 |

### Linguistic Analysis

Cohesive features in all 960 essays were measured using the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, 2016). TAACO is a free automated natural language processing tool ().

TAACO incorporates over 150 classic and recently developed text cohesion measures related to local, global, and textual cohesion which would be impractical to calculate by hand. It provides direct, automatic measurement of cohesive features in a text including local cohesive features that connect words and sentences. Such indices include the use of lexical connectives (because, and) and lexical overlap between sentences. Global cohesive features examine links between larger text segments (e.g., lexical overlap across paragraphs). Textual cohesion measurements include elements of cohesion that apply to the scope of an entire text, such as discourse orientation features. These may be type-token ratio (TTR) or text-wide repetition of a certain cohesive feature (e.g., referential pronouns). In previous studies, TAACO indices have been found to significantly correlate with and predict human judgements of coherence and overall text quality in SAT independent essay writing (Crossley, Kyle & McNamara, 2016). The program has since been used to identify and calculate cohesion variables in models of second language writing development (Kim & Crossley, 2018).

### Local and Global Overlap Indices

Lexical overlap as a measure of cohesion can be traced back to Hoey's (1991) assertion that repetition is a common strategy for explicit marking of cohesion. TAACO calculates a number of overlap indices that measure the degree to which words or lemmas are repeated in subsequent text sections: sentences (locally) and paragraphs (globally). Overlap by lemma is computed between two adjacent sentences or paragraphs or three adjacent sentences or paragraphs. TAACO performs separate calculations for all lemma overlap, for content word lemma overlap and function word lemma overlap, as well as separate calculations for lemma overlap for parts of speech such as nouns, verbs, adjectives, adverbs, and pronouns. It also provides lemma overlap calculations normed for number of sentences and paragraphs, and as binary overlap measurement (i.e., whether there is any overlap between adjacent sentences or paragraphs).

Local cohesion indices have demonstrated positive relations with other measures of cohesion such as coreference and type-token ratio. (McNamara et al., 2010), but generally they demonstrate no significant relations with measures of coherence (Crossley & McNamara, 2011a, 2011b). However, global overlap indices have demonstrated positive relations with measures of text coherence in previous studies (Crossley &McNamara, 2011b; Crossley et al., 2016).

### Connectives

TAACO also measures the incidence of connectives. Many of the connective indices are similar to those found in Coh-Metrix (McNamara et al., 2014). Theoretically motivated connective indices include positive and negative connectives, and classes of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001), such as logical, additive, and causative connectives. TAACO also reports on connectives related to rhetorical devices including disjunctions and contrasting connectives. Indices measuring the use of connectives have previously demonstrated a negative, if any, correlation with essay quality and essay coherence (Crossley & McNamara, 2011a, 2011b), although some connectives have demonstrated positive relations with other cohesion measures (McNamara et al., 2010).

### Textual Cohesion Indices

TAACO calculates cohesion across an entire text through measurements of givenness and type-token ratio (TTR). Like the local and global cohesion measurements, textual cohesion measurements calculate an average score based on specific incidences of a linguistic construct. Unlike the local and global cohesion measurements, textual cohesion measurements are not based on a specific locality of coreference but count aspects of cohesion that occur throughout a text, not restricted to discrete localities. Givenness refers to the amount of information that is recoverable from the preceding discourse. To assess givenness, TAACO calculates the incidence of a variety of pronoun types under the premise that pronouns are used when information is given (Crossley, Allen, Kyle, & McNamara, 2014). Pronouns in TAACO include first-person, third-person pronouns, subject pronouns, and quantity pronouns. Similarly, TAACO calculates the ratio of nouns to pronouns under the presumption that lower ratios imply higher levels of givenness. TAACO also calculates the incidence of definite articles (i.e., the) and demonstratives (i.e., this, those, that, and these) found in a text, under the presumption that the definiteness of these words belies given information. Likewise, the number and proportion of single content lemmas (e.g., how many lemmas occur only once in a text) are also counted. Givenness indices have previously been shown to relate positively with measures of text coherence (Crossley & McNamara, 2011b, 2011c).

TAACO calculates a number of different TTR indices which measure the inverse of repetition of words in the text by dividing the number of individual words (types) by the total number of words (tokens). TTR indices calculated by TAACO include overall TTR, lemmatized TTR, and TTR indices specifically for content and function words as well as bigrams and trigrams. TTR indices have demonstrated inverse relations with measures of cohesion and text coherence (Crossley & McNamara, 2014; McNamara et al., 2010). However, it is not a direct measurement of cohesion as TTR is on its surface a measure of lexical diversity and an inverse of cohesion (McCarthy & Jarvis, 2010), although it is flawed due to its strong relationship with text length in words (which is not related to cohesion). However, calculating TTR by lemmas (e.g. need and needed would be considered as two tokens of the same type) somewhat normalizes TTR and decreases the likelihood of multicollinearity with text length.

Each index in TAACO was calculated for each text. As each measure included has some basis as an operationalization of cohesion, we do not exclude any indices from analysis on concept alone, instead taking a more data driven approach. For more information about how indices were derived and how they are specifically related to cohesion, beyond what is described in the current study (see Crossley, Kyle, & McNamara, 2016b).

### Statistical Analysis

Prior to any analysis, the corpus was organized by task and prompt so that a training set and test set which accounted for the repeated measures nature of the data along with prompt differences could be created. The first test administration's integrated essays (n = 240) and the second test administration's independent essays (n = 240) were used in the training in order ensure independence of the data (i.e., avoid repeated sampling) as well as develop models based on separate prompts. The model built from the first split of the essays was then applied to the second split (the remaining 240 independent and 240 integrated essays) which acted as a test set.

In order to determine the relationship between cohesion features and L2 independent and integrated writing tasks, a number of statistical analyses were conducted. We first conducted a Multivariate Analysis of Variance (MANOVA) and post-hoc pair-wise tests to identify which cohesive features were significantly different between test tasks in the training set. This was followed by a Discriminant Function Analysis (DFA) to identify features that classified writing tasks in order to better understand the predictive power of cohesive indices. This analysis was followed by a linear regression model to examine if cohesion devices which could predict test task in the DFA were also predictive of human judgments of essay quality.

Cohesion features used in this analysis were first assessed to ensure they were not strongly correlated with number of words in a text ($r > .7$) because the number of words alone is a strong indicator of differences in writing tasks as well strong predictors of human judgments (Crossley & McNamara, 2012). Text length varied widely in the TOEFL corpus, so if an index was strongly correlated with text length, it was not included in the analysis. Eliminating features strongly correlated with text length allows us to examine the effect of cohesive features beyond text length without needing to trim the data around an arbitrary text length.

Cohesion features were also examined for normality of distribution and for multicollinearity with each other ($r > .900$). If two or more features were multicollinear, all but one was removed. After checking for these criteria, a multivariate analysis of variance (MANOVA) was conducted to indicate which features were significantly different between the independent and integrated writing samples in the training set. To account for multiple comparisons, only those indices which demonstrated a significant difference between independent and integrated writing with $p < .001$ were included. Setting such a low threshold for critical values can offset the chance that significance was found merely due to repeated pair-wise comparison in the post-hoc analyses. The independent variable in this analysis was writing task, and the dependent variables were cohesive features. Measurements from the training set indices which were significant in the MANOVA were then entered into a DFA model. The DFA was carried out using R (R Core Team, 2007). DFA allows group classification based on a number of predictors (i.e., membership as either an independent or integrated essay). In DFA modelling, if variables demonstrated suppression effect, the variable was removed from consideration to ensure that the final model reflected the initial statistical trends (Jarvis, 2011; Kyle & Crossley, 2016). Suppression effects occur when an independent variable, which in isolation is associated with a dependent variable, appears in multivariate modeling with a coefficient inverse to its expected direction of association as an optimal model is constructed.

In order to analyze the connection between cohesive features and human judgments of L2 writing quality, linear regression models were next developed using R. We developed two regression models for each test task using variables found to be predictive of task type in the DFA model described above. The first examined links between cohesion features and independent essay quality. The second examined links between cohesion features and integrated essay quality. The same training and test set

split was maintained for the linear regressions but separated by task for these analyses. In both models, we controlled for suppression effects, and only indices which had a significant ($p < .001$) and meaningful ($r > .1$) partial correlation with score in the respective task were included in the analysis.

## Results

### Results from Test Task MANOVA

After checking for multicollinearity, normality, and correlation with text length within the training set, 29 indices were included in the MANOVA model. Of these, 24 variables demonstrated significant differences between writing tasks with $p < .001$, a level set to control for multiple comparisons. Descriptive statistics for these 24 indices are given in Appendix A. These variables indicated that integrated and independent writing are significantly different across levels of cohesion: local cohesion variables (e.g., lemma overlap across two or three adjacent sentences), use of connectives, subordinators and determiners, global cohesion variables (e.g., overlap of personal pronouns across two or three adjacent paragraphs), and textual cohesion (e.g., span of coreference chains, relevance of sentences and paragraphs to the overall text). Specifically, integrated writing tasks contained more lexical overlap on the local level and generally had more repeated nouns and pronouns. Integrated writing was marked by more use of determiners (e.g., "the" and "that"), subordinators (e.g., "although" and "as"), and negative logical connectors (e.g., "however"). Independent writing tasks contained more global cohesive features, such as lexical overlap across adjacent paragraphs (especially of verbs and function words), and textual cohesive features (such as length of coreference chains and average relevance of sentences and paragraphs to the overall topic). Independent writing was marked by use of specific cohesive connectors of addition (e.g., "also" and "additionally") and logical connection (e.g., "therefore"). MANOVA results are presented in Appendix B.

### Discriminant Function Analysis of Task Type

#### *Training Set*

The 24 indices determined to be significantly different between the two text types at the $p < .005$ level (indicated in Appendix B) were entered into a Discriminant Function Analysis (DFA) classification model, which can provide evidence that these indices discriminate between the two task types. The DFA was validated using Ten-fold cross-validation, whereby a model using the selected indices is constructed based on 90% of our observations (i.e., texts) and compared to the remaining 10% to predict task type using the most predictive indices. This process is repeated ten times, each time with a different 10% left out for comparison. Six cohesion indices were ultimately retained in the model as significant predictors of task type in the training set. The six significant predictors are listed in order of predictive power in Table 2. The results show that the DFA using the six predictive indices correctly classified 450 of the 480 texts as either integrated or independent, $\chi 2\ (1, 480) = 367.5$, $p < .001$ with 93.6% accuracy, compared to the baseline chance of 50.0%. The reported Kappa of .873, indicating very strong agreement between the actual test task and the predicted task classification. The confusion matrix for the DFA training set is shown in Table 3. The results indicate that, in the TOEFL iBT, independent writing is marked by greater use of repeated pronouns, use of logical connectives, and overlap of function words across paragraphs, while integrated writing is associated with greater use of determiners, content word overlap across three adjacent sentences, and a higher type-token ratio

#### *Test Set*

The DFA model from the training set was also applied, again using ten-fold cross-validation, to the test set. The results in the test set were similar to the training set results, and the model correctly

classified essays as being independent or integrated in 92.3% of cases in the test set, which was significantly higher than the baseline accuracy of 50.00%, $\chi 2$ (1, 480) = 343.70, p < .001). The reported Kappa = .846 indicates strong agreement between actual and predicted test task. The confusion matrix for the DFA test set is also shown in Table 3.

**Table 2** *Discriminant Function Analysis predictors generated in the training set and used in the test set*

| Loading | Index | Text type association | Level of cohesion |
|---|---|---|---|
| **1** | Number of determiners | Integrated | Text |
| **2** | Repeated use of pronouns throughout the text | Independent | Text |
| **3** | Function word lemma overlap across 3 adjacent paragraphs per sentence | Independent | Global |
| **4** | Lemma Type-token ratio | Integrated | Text |
| **5** | Content word overlap across three adjacent sentences | Integrated | Local |
| **6** | Number of logical connectives | Independent | Local |

**Table 3** *Confusion Matrix of task type predictions*

| Prediction | Membership | |
|---|---|---|
| | Integrated | Independent |
| **Training** | | |
| **Integrated** | 225 | 15 |
| **Independent** | 15 | 225 |
| **Test** | | |
| **Integrated** | 225 | 22 |
| **Independent** | 15 | 218 |

### Correlations Training Set: Integrated Writing

Partial correlations (correlations between a predictor and dependent variable controlling for other predictors) were conducted between the six cohesion indices which showed predictive power in the DFA model and the human ratings of writing quality for the 240 integrated essays selected for the training set. Two indices demonstrated significant correlation with at least a small effect size ($p < .01$, $r > .10$ or $r < -.10$) with the human ratings and had a moderate effect size. See Table 4 for a list of variables with significant correlations.

**Table 4** *Selected indices for regression analysis of the integrated essays: Training set.*

| Index | Category | partial *r* | *p* |
|---|---|---|---|
| Function word lemma overlap across three adjacent paragraphs per sentence | Global Cohesion | 0.250 | <.001 |
| Lemma type-token ratio | Text Cohesion | -0.210 | <.001 |

### Integrated Writing Regression Analysis

A stepwise regression analysis to predict human ratings of essay quality using the cohesion indices

found in the DFA models yielded a significant model, $F$ (2, 237) = 23.550, $p < .001$, $r = .399$, $r^2 = .159$. Two indices were included as significant predictors of the human ratings: *Type-token ratio by lemma* and *Function word overlap across three adjacent paragraphs*. *Type-token ratio by lemma* is negatively correlated with the scores, and *Function word overlap across three adjacent paragraphs* is positively correlated with the scores, indicating that essays with lower lemma type-token ratio and more global overlap of function words received higher scores. See Table 5 for the model built from the training set.

**Table 5** *Regression analysis findings to predict the integrated essay scores: Training set.*

| Entry | Index | *r* | $r^2$ | *B* | B | S.E. |
|---|---|---|---|---|---|---|
| **Entry 1** | Function word lemma overlap across three adjacent paragraphs per sentence | 0.359 | 0.129 | 0.110 | 0.274 | 0.027 |
| **Entry 2** | Lemma type-token ratio | 0.399 | 0.159 | -4.11 | -0.203 | 1.344 |

Notes: *B* = unstandardized $\beta$; B = standardized $\beta$; S.E. = standard error. Estimated constant term is 2.282.

We used the model reported in the training set to predict the human scores in the test set. To determine the predictive power of the two indices retained in the regression model, we computed an estimated score for each integrated essay in the 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 together 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 $r = 0.290$, $r^2 = 0.084$. The results from the test set model demonstrated that the combination of the two predictors accounted for 8.4% of the variance in the assigned scores of the 240 integrated essays in the test set, providing some confidence for the generalizability of our model for integrated essays using indices from the text classification DFA.

### Correlations Training Set: Independent Writing

Partial correlations were conducted between the six cohesion indices which showed predictive power in the DFA model and the human ratings of writing quality for the 240 independent essays selected for the training set. Four indices demonstrated significant correlation with at least a small effect size ($p < .01$, $r > .10$ or $r < -.10$) with the human ratings, with only one having a moderate effect size. See Table 6 for a list of variables with significant correlations.

**Table 6** *Selected indices for regression analysis of the independent essays: Training set*

| Index | Category | partial *r* | *p* |
|---|---|---|---|
| Function word lemma overlap across three adjacent paragraphs per sentence | Global Cohesion | 0.298 | <.001 |
| Lemma type-token ratio | Text Cohesion | -0.166 | <.001 |
| Use of determiners throughout the text | Text Cohesion | 0.153 | <.001 |
| Use of logical connectives throughout the text | Local Cohesion | -0.204 | <.001 |

### Independent Writing Regression Analysis

A stepwise regression analysis using as independent variables the four indices which were significant

predictors of text type to predict human ratings of writing quality in independent essays yielded a significant model, $F$ (4, 235) = 17.300, $p < .001$ $r = .463$, $r^2 = .214$. All four indices were included as significant predictors of the human ratings, including the two appearing in the integrated essay model: *Function word overlap across three adjacent paragraphs, use of logical connectives, use of determiners*, and *lemma type-token ratio*. As with integrated writing, *function word overlap across three adjacent paragraphs* is positively correlated with human ratings, and *lemma type-token ratio* is negatively correlated. In this model, *use of determiners* is positively correlated with human ratings, and *use of logical connectives* is negatively correlated. This indicates that human ratings are predicted by higher use of determiners and lower use of logical connectives, in addition to having high global overlap of function words and low lemma type-token ratio. The model demonstrated that the three indices explained 21.4% of the variance in the human ratings of essay quality in independent writing. See Table 7 for additional information.

**Table 7** *Regression analysis findings to predict independent essay scores: Training set.*

| Entry | Index | *r* | $r^2$ | *B* | B | S.E. |
|---|---|---|---|---|---|---|
| **Entry 1** | Function word lemma overlap across three adjacent paragraphs per sentence | 0.396 | 0.157 | 0.063 | 0.310 | 0.013 |
| **Entry 2** | Use of Logical Connectives throughout the text | 0.424 | 0.180 | -10.793 | -0.198 | 3.378 |
| **Entry 3** | Use of determiners throughout the text | 0.446 | 0.199 | 4.600 | 0.140 | 1.943 |
| **Entry 4** | Lemma type-token ratio | 0.463 | 0.214 | -2.32 | -0.175 | 0.901 |

Notes: $r^2$ = adjusted r squared; $B$ = unstandardized $\beta$; B = standardized $\beta$; S.E. = standard error. Estimated constant term is 3.890.

We used the model reported in the training set to predict the human scores in the test set in the same fashion as with the integrated essays. The regression model, when applied to the test set, reported $r$ = 0.410, $r^2$ = 0.168. The results from the test set model demonstrated that the combination of the four predictors accounted for 16.8% of the variance in the assigned scores of the 240 integrated essays in the test set, providing confidence in the generalizability of our model for independent essay using indices from the text classification DFA.

## Discussion

In this study, features of cohesion in a learner writing corpus derived from a standardized assessment were used to classify learner texts according to writing task and to predict writing quality using NLP tools and statistical modeling. Results indicate that an accurate model for predicting writing assessment task type can be constructed using a number of linguistic features. Additionally, a subset of these features were important predictors of human judgments of writing quality, explaining a significant portion of the variance in both L2 source-based and independent writing. This reflects the notion that while cohesive features may vary by text type, they are an important element of L2 writing quality.

In showing that the tasks in the TOEFL (i.e., integrated writing and independent writing) can be automatically classified using a Discriminant Function Analysis based on cohesion indices with 92.3% precision, this study finds support for the long-held assertions that surface features vary significantly between text types (Biber, 1988) and that cohesive features vary by text type (Halliday & Hasan, 1976). These differences between integrated and independent writing parallel the distinction which Biber (1988) found between informational and interactive texts and highlight the interference we may see from the topics. The integrated tasks were both on scientific topics and are framed in a way that push the writers to take a more informational stance, where the independent tasks were both on more

personal matters and may have led to a more interactional stance. This overlap between the writing task genres in the TOEFL and the established informational/interactive dimension in many register analyses is worth further investigation.

Specifically, independent writing was found to have more repeated use of pronouns throughout a text and more use of logical connectives. The purpose of independent writing (i.e., persuasive writing) may lead writers to produce a great number of logical connectives to independently frame their arguments. Additionally, as the writers have less linguistic resources given by the task prompt, writing is more centered around a core topic, leading to the use of more pronouns. These features are highlighted in examples (1) and (2), with logical connectives in bold and pronoun use in italics.

(1) **Therefore**, working together with other partners on a project, **for instance**, is vital to *its* successful completion.
(2) My parents, **for instance**, cannot work with computer. **However**, *they* gradually begin to need *this* at *their* working place **and** *they* rely on me to teach *them*. **On the other hand**, there were no computers and such complicated technologies in the past.

Independent essays were also found to contain more function word overlap across paragraphs than in integrated writing, again due in part to the independent writing task giving writers fewer linguistic resources to work from. This aspect is harder to exemplify as it includes many different surface features and spans, oftentimes, the length of a text. It is firstly dependent on a writer's ability to organize their text by paragraph and competence in using function word categories. More related to cohesion and discourse orientation, function word overlap between paragraphs shows, to some extent, a common structuring of paragraphs within a text, as the function word categories provide framing for discourse. Thus, for independent writing, if one paragraph involves introducing a main idea, providing a detail, and then linking to a counterpoint with however, a subsequent paragraph may likely follow the same discursive signposting.

In relation to integrated writing, this study finds that integrated writing on the TOEFL involves more local cohesion in the form of content word overlap across any three adjacent sentences and greater use of determiners as exemplified in examples (3) and (4). Determiners are marked in bold, and content word overlap is marked with italics.

(3) In **the** lecture *points* were raised that explained **the** limitations of **the three** theories in **the** reading passage. **The** *points* mentioned are…
(4) Firstly, *birds migrate* even when the *sun* and *stars* are not visible. As result **the** theory of celestial navigation is completely not *true*; it is partially *true*. It is *true* that *birds* use **the** *sun* and *stars* to *migrate*…

Integrated writing on the TOEFL includes a source to which writers must respond so there is increased lexical resources to draw upon and a narrower focus in topic which can explain the lexical overlap of content words. The fact that the integrated writing is a response also helps explain why there is more determiner use, as much of the content in the texts is marked for givenness. The linguistic resources of the integrated writing source provide writers a set of cues for what words are rich in topic relevance, providing motivation for the repetition of content words and the greater use of determiners (which indicate given information).

Although it may seem contrary to the above examples, TTR is also higher for integrated writing, likely stemming from what integrated writing lacks compared to independent writing: pronoun repetition and

function word overlap. Writers writing from a source have access to the linguistic resources of a text as well as their own pre-existing linguistic resources, which may lead to the higher type-token ratio we find in integrated writing. As there are less repetitions of function word types, and again more linguistic resources to draw upon in the source text, TTR raises significantly for integrated texts.

When the DFA was unable to correctly classify a text as either integrated or independent, there were technical reasons for the mismatch in some cases. Independent texts were associated with function word overlap across multiple paragraphs. An independent text that did not have clearly marked paragraphs would be more likely to be classified as Integrated as the overlap score would be registered as zero (0). Additionally, as Integrated texts were associated with higher use of determiners, repeated errors with irregular determiner use (e.g., using "sun" instead of "the sun") made it more likely an Integrated text would be classified as an Independent text.

Regarding cohesion and writing quality, the models for predicting writing quality were significant, but not as accurate as the task classification model. This is expected, because cohesion is only one component of the quality in L2 writing. In the absence of other linguistic features (e.g., lexical sophistication, syntactic complexity), the results in this study show cohesion indices account for a reasonable amount variance in L2 writing quality. However, in addition to the similarity of explained score variance in models for both tasks, the cohesion features that explained variance were similar in each model, indicating that the aspects of cohesion used in human judgments of writing quality are not different between assessment tasks.

For independent writing, the model predicted 17% of the variance and included four predictors. Two of these were strongly associated with independent writing in the DFA (function word overlap between paragraphs and logical connectives), but two indices were more strongly associated with integrated writing (type-token ratio and use of determiners). Additionally, while the incidence of logical connectives was a positive predictor of independent writing in the DFA, it was negatively correlated with the quality in the independent set. Conversely, the use of determiners was positively associated with integrated writing in the DFA but was positively correlated with independent writing score. For TOEFL integrated writing, function word overlap across paragraphs and type-token ratio were the only cohesion indices predictive of score, but only type-token ratio was associated positively with integrated writing in the DFA. In sum, a linguistic feature associated with a particular type of writing is not necessarily seen as positive for that type of writing.

The distinction between what L2 writers actually do in each task (as shown in the DFA) and what is predictive or not predictive of the quality (as shown in the regression models) has precedent in L2 writing literature. For instance, coherence is not bound to the explicit use of cohesive devices (Carrell, 1982; Crewe, 1990; Plakans & Gebril, 2017; Williams, 2012) and with high knowledge readers such as expert raters, the use of explicit cohesive devices may lead to a less coherent text (McNamara et al., 1996, O'Reilly & McNamara, 2007). Such findings have been reported in previous L2 studies (Crossley, Kyle & McNamara, 2016a, Guo et al., 2013).

L2 writers may produce a greater number of logical connectives because they believe that the use of logical connectors makes their writing more sophisticated, especially in non-communicative, high-stakes formats (Williams, 2012). L2 writers receive explicit and precise instruction on connective use, possibly leading to overuse (Crewe, 1990). Increased used of connectives in independent writing may be task related as well because writers depend on them as road-mapping devices to construct meaningful arguments (Spivey, 1990; Yang, 2014). Beyond logical connectives, the incidence of determiners was strongly related to integrated writing but was a strong predictor of writing quality for independent writing. Determiners, such as articles, are difficult for L2 learners to master (Master,

2002), and unlike integrated writing where the source may provide guidance for the use of determiners, independent essays likely better demonstrate a writer's mastery of difficult linguistic features. Thus, writers in the independent task that produced fewer determiners may have been viewed by expert raters as less proficient.

In looking at the features which were predictive of the quality for both tasks, we find that type-token ratio was included in both models. This affirms results in previous studies which have reported type-token ratio as an important predictor of writing quality (Guo et al., 2013; McNamara et al, 2010). Function word overlap across adjacent paragraphs was also a significant predictor in both models. This stands to reason since use of function words, such as prepositions and determiners, are difficult to master for L2 English learners (Crossley, Kyle, & McNamara, 2016a; De Felice & Pulman, 2008) and can be seen as a general feature of more sophisticated writing. Additionally, global cohesion as reported by paragraph overlap measures have also been associated with greater writing quality (Crossley, Kyle, & McNamara, 2016b; Guo et al., 2013). Specific to function words, overlap across paragraphs may be indicative of a coherent structure and organization across paragraphs. Examples (5) and (6) provide brief excerpts from a single lower-scoring independent writing.

(5) If we don't have any memoriziable [sic] memories what is for living? If you don't have anyting [sic] to tell others that you loved sharing it with others. Just for living another's words to take a breath? Is that it ? No , that is not that.
(6) I mean we are busy. We are always in a rush , always have a lot things going on with us , and we can managed almost whatever we want.

These excerpts are from the same text but different paragraphs. The author attempts a parallel structure sentence by sentence, but there is very little overlap between paragraphs in how these pieces of discourse are hung together.

Conversely, in (7) and (8) below, we see two excerpts from different paragraphs within one highly rated integrated writing sample. Function word overlaps between the excerpts from paragraphs are bolded, and repetitions in general are italicized. In addition to repeated use of function words throughout both parts of the text, there are many repetitions of words in general, indicative of this text's overall low type token ratio.

(7) ***this*** *theory* means ***that*** if a *bird* losts in a *place* ***that*** unfamiliar with them ***they*** couldn't go back or find a way out from ***that*** *place*. it's ***because they don't*** *have* any memory ***about that*** *place*, ***they*** *haven't* been there.
(8) … ***this*** *theory* ***doesn't*** explain all the things *needed* by *birds* to navigate. ***because*** when *birds* fly, ***they do not*** *need* only directions, ***they*** also *need* information ***about*** how far…

Here, we can see how the author is commenting on different subtopics in their integrated essay but maintain a similar approach to sequencing ideas. They expand on a point with a phrase like "this theory means" or "this theory doesn't explain" and connect ideas logically in both paragraphs using "because." Seeing the use of repeated function words in this way in highly rated texts provides some explanation for why function word lemma overlap across paragraphs and low lemma type-token ratio predicted higher scoring texts.

The mismatch between the features that distinguish tasks and how those features predict essay quality is not, in itself, problematic. Rather, the results show that judgements of writing quality are predicated

on similar features of cohesion across tasks. However, the tasks themselves seem to lead to the production of different types of cohesion feature. This difference warrants the use of multiple writing tasks to elicit different facets of academic writing in L2 writing assessment and raises the questions of whether the assessment of cohesion requires further definition in terms of the register of writing in standardized assessment rubrics beyond their mention under organization and coherence (as in the TOEFL rubrics). Even in assessments where cohesion is separately assessed with an analytic rubric (as in the IELTS rubrics), cohesion is not operationalized differently for the more integrated descriptive task and the prompted essay task. To capitalize on measuring academic writing using multiple writing tasks, rating tools should be built around the specific register of the writing task with specific attention to the unique features of the register that create a text's texture. Alternatively, it may be the case that source texts need to be tightly controlled to ensure minimal linguistic priming.

## Conclusion

### Summary of the Findings

The results of this study show not only the importance of examining a wide and varied set of cohesive features to gain a fine-grained understanding of L2 writing quality, pinpointing exact features relevant to writing score variance but also the importance of understanding how task type can influence the production of cohesion features. The interaction between task types and writing quality has important implications for L2 writing for both the development of standardized tests and the assessment of writing within these tests. The findings also have implications for the automatic assessment of essay quality using cohesion features.

For assessment developers and practitioners, this study highlights the interaction between cohesion and text type. As cohesive features include strategies writers employ for organizing ideas, and test tasks are bound to elicit a specific set of writing strategies, different tasks will likely evoke different sets of cohesive features. For writing assessment, it is important to recognize that different writing tasks may not elicit cohesive strategies that are associated with quality writing.

### Limitations

This study has several limitations. The issue of prompt effects provides a large hurdle for this study. Especially for analysis of integrated writing assessment, where prompts affect a wide range of linguistic resources available to the examinees, prompt will likely confound some the results found regarding prevalence of cohesive features and this was not addressed here. Additionally, the explicit overlap between source texts in integrated writing and the examinees' writing was not analyzed, and this may be a further compounding factor.

The findings from this study highlighted specifically the extent to which cohesive indices which distinguish writing assessment tasks also predict writing quality. As such, the results from this study do cannot put forth a unified description of how cohesive features affect writing quality and the results may not be generalizable for all writing contexts. Related to this, the data-driven nature of the study means indices of cohesion which may be highly predictive of writing quality but did not show initial significant difference in use between writing tasks would not become apparent in this study. Although this study was not able to produce a specific set of features which both typified writing tasks and uniquely predicted individual task scores, the results show the benefit of exploring cohesion through model-building while controlling for task differences. The derived models can help test designers improve writing assessment in the classroom and in standardized testing. The models may also be applied in automatic text classification tasks as well as automatic essay scoring tasks.

### Future Directions

Future studies should consider additional cohesion features such as register-specific features and inter-textual cohesion features that measure overlap between documents (e.g., source and texts). Newer lexical diversity indices that control for text length or include features related to richness, disparity, dispersion, and evenness (Jarvis, 2002; 2013; McCarthy et al., 2010) may also provide additional evidence for the role of cohesion features in independent and integrated writing samples. Lastly, while difficult to measure computationally, the accurate use of cohesion features may provide greater details about how they interact in L2 writing. Future research into fine-grained indices of cohesion and L2 writing should expand into a broader range of task types to identify cohesive features which can predict quality for those tasks.

## References

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL iBT® Test: A lexico-grammatical analysis. *ETS Research Report Series*, *2013*(1), i-128.

Bunch, G. C., & Willett, K. (2013). Writing to mean in middle school: Understanding how second language writers negotiate textually-rich content-area instruction. *Journal of Second Language Writing, 22*(2), 141−160. https://doi.org/10.1016/j.jslw.2013.03.007

British Council. (2018). IELTS assessment criteria. Retrieved December 26, 2018, from https://takeielts.britishcouncil.org/find-out-about-results/ielts-assessment-criteria

Carrell, P. (1982). Cohesion is not coherence. *TESOL Quarterly*, *16*, 479-488.

Celce-Murcia, M., & Olshtain, E. (2000). *Discourse and context in language teaching: A guide for language teachers*. New York, NY: Cambridge University Press.

Crewe, W. J. (1990). The illogic of logical connectives. *ELT Journal, 44*(4), 316-25.

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

Crossley, S. A., Dascalu, M., Trausan-Matu, S., Allen, L., & McNamara, D. (2016). Document cohesion flow: Striving towards coherence. *38th Annual Meeting of the Cognitive Science Society*, pp. 764-769. Cognitive Science Society, Philadelphia

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1-16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., & McNamara, D. S. (2011a). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, *21*(2-3), 170-191.

Crossley, S. A., & McNamara, D. S. (2011b). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society.* (pp. 1236-1241). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011c). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing, 20,* 271-285. doi:10.1016/j.jslw.2011.05.007

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*(2), 115-135.

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.

Dascalu, Mihai, Trausan-Matu, S., McNamara, D., & Dessus, P. (2015). ReaderBench: Automated evaluation of collaboration based on cohesion and dialogism. *International Journal of Computer-Supported Collaborative Learning, 10*(4), 395-423. https://doi.org/10.1007/s11412-015-9226-y

De Felice, R., & Pulman, S. G. (2008). A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics* (Coling 2008) (pp. 169-176). Manchester, UK: Coling 2008 Organizing Committee. Retrieved from http://www.aclweb.org/anthology/C08-1022

De Silva, R. (2015). Writing strategy instruction: Its impact on writing in a second language for academic purposes. *Language Teaching Research, 19*(3), 301-323. https://doi.org/10.1177/1362168814541738

Donaldson, M. L., Reid, J., & Murray, C. (2007). Causal sentence production in children with language impairments. *International Journal of Language and Communication Disorders, 42*(2), 155-186.

Duggleby, S. J., Tang, W., & Kuo-Newhouse, A. (2016). Does the use of connective words in written assessments predict high school students' reading and writing achievement? *Reading Psychology, 37*(4), 511-532.

Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods, 39,* 212-223.

Duran, N. D., Hall, C., McCarthy, P. M., & McNamara, D. S. (2010). The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics, 31,* 439-462.

Educational Testing Service. (2012). The official guide to the TOEFL test (4th ed.). New York, NY: Educational Testing Service.

Ekiert, M. & Han, Z. (2016). L1-fraught difficulty: The case of L2 acquisition of English. In R. Alonso Alonso (Ed.), *Crosslinguistic influence in second language acquisition*. Bristol: Multilingual Matters.

Foltz, P. W. (2007). Discourse coherence and LSA. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 167-184). Mahwah, NJ: Erlbaum.

Friginal, E., Li, M., & Weigle, S. C. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing, 23,* 1-16. https://doi.org/10.1016/j.jslw.2013.10.001

Galloway, E., & Uccelli, P. (2015). Modeling the relationship between lexico-grammatical and discourse organization skills in middle grade writers: insights into later productive language skills that support academic writing. *Reading & Writing, 28*(6), 797-828. https://doi.org/10.1007/s11145-015-9550-7

García, J.-N., & Fidalgo, R. (2008). Orchestration of writing processes and writing products: A comparison of sixth-grade students with and without learning disabilities. *Learning Disabilities: A Contemporary Journal, 6*(2), 77-98.

Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialog. *Information Design Journal, 15,* 199-213.

Grabe, W. (1987). Contrastive rhetoric and text-type research. In U. Connor & R. Kaplan (Eds.), *Writing across languages: Analysis of L2 text* (pp. 115-138). Reading, MA: Addison-Wesley.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective.* Longman: New York.

Grabe, W., & Zhang, C. (2016). Focus on texts and readers: Linguistic and rhetorical features. In R. M. Manchon & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 178-192). Boston: DeGruyter.

Grice, H. P. (1975). Logic and conversation. In Cole, P. & Morgan, J. L. (Eds.) *Syntax and semantics Vol 3 (speech acts)* (pp. 41-58). New York, NY: Academic Press,

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. *Assessing Writing, 18,* 218-238. https://doi.org/10.1016/j.asw.2013.05.002

Hall, S. S., Kowalski, R., Paterson, K. B., Basran, J., Filik, R., & Maltby, J. (2015). Local text cohesion, reading ability and individual science aspirations: Key factors influencing comprehension in science classes. *British Educational Research Journal, 41*(1), 122-142. https://doi.org/10.1002/berj.3134

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

Hempelmann, C. F., Dufty, D., McCarthy, P. M., Graesser, A. C., Cai, Z., & McNamara, D. S. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 941-946). Mahwah, NJ: Erlbaum.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Horning, A. S., Kraemer, E. W., & WAC Clearinghouse (Firm). (2013). *Reconnecting reading and writing*. Anderson, S.C.: Parlor Press and the WAC Clearinghouse.

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

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87-106.

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing, 34*(4), 537-553.

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37*, 39-56. https://doi.org/10.1016/j.asw.2018.03.002

Koutsoftas, A. D., & Petersen, V. (2017). Written cohesion in children with and without language learning disabilities. *International Journal of Language & Communication Disorders, 52*(5), 612-625. https://doi.org/10.1111/1460-6984.12306

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34,* 12-24. doi:10.1016/j.jslw.2016.10.003

Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System, 33*(4), 623-636.

Lux, P., & Grabe, W. (1991). Multivariate approaches to contrastive rhetoric. *Lenguas Modernas, 18,* 133-160.

McCarthy, M., & Carter, R. (1994). *Language as discourse: Perspectives for language teaching.* New York: Longman.

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, 42*(2), 381-392.

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

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86. https://doi.org/10.1177/0741088309351547

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

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292-330.

Mullis, I. V. S., & Mellon, J. C. (1980). *Guidelines for describing three aspects of writing: Syntax, cohesion and mechanics*. Retrieved from https://eric.ed.gov/?id=ED205572

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

Pinto, G., Tarchi, C., & Bigozzi, L. (2015). The relationship between oral and written narratives: A three-year longitudinal study of narrative cohesion, coherence, and structure. *British Journal of Educational Psychology, 85*(4), 551-569.

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing, 31*(1), 98-112. https://doi.org/10.1016/j.asw.2016.08.005

Polat, M. (2015). Developing a writing assessment profile. *Proceedings of the Multidisciplinary Academic Conference*, 1-12.

Reed, D. K., & Kershaw-Herrera, S. (2016). An examination of text complexity as characterized by readability and cohesion. *Journal of Experimental Education, 84*(1), 75-97. https://doi.org/10.1080/00220973.2014.963214

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*, pp. 79-107.

Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional perspective*. (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff.

Reynolds, D. W. (2001). Language in the balance: Lexical repetition as a function of topic, cultural background, and writing development. *Language Learning, 51*(3), 437-476.

Reynolds, G. A., & Perin, D., (2009). A comparison of text structure and self-regulated writing strategies for composing from sources by middle school students. *Reading Psychology, 30*(3), 265-300. https://doi.org/10.1080/02702710802411547

Román, D. X., Briceño, A., Rohde, H., & Hironaka, S. (2016). Linguistic cohesion in middle-school texts: A comparison of logical connectives usage in science and social studies textbooks. *Electronic Journal of Science Education, 20*(6), 1-19.

Ruegg, R., & Sugiyama, Y. (2013). Organization of ideas in writing: What are raters sensitive to? *Language Testing in Asia, 3*(1), 8. https://doi.org/10.1186/2229-0443-3-8

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing, 9*(3), 259-291. https://doi.org/10.1016/S1060-3743(00)00028-X

Scholes, R., & Comley, N. (1985). *The practice of writing.* New York: St. Martin's Press.

Shaw, S. & Weir, C. (2007). *Examining writing: Research and practice in assessing second language writing* (Vol. 26). Cambridge: Cambridge University Press.

Sheehan, K. (2013). Measuring cohesion: An approach that accounts for differences in the degree of integration challenge presented by different types of sentences. *Educational Measurement: Issues and Practice, 32*(4), 28-37.

Silva, T. (2016). An overview of the development of the infrastructure of second language writing studies. In R. M. Manchon & P. K. Matsuda (Eds.), *Handbook of Second and Foreign Language Writing* (pp. 178-192). Boston: DeGruyter.

Spivey, N. (1990). Transforming texts: Constructive processes in reading and writing. *Written Communication, 7*(2), 256-287.

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing, 32,* 17-35. doi:10.1016/j.jslw.2016.02.002

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills*. Ann Arbor, Mich.: University of Michigan Press.

Watson Todd, R., Khongput, S., & Darasawang, P. (2007). Coherence, cohesion and comments on students' academic essays. *Assessing Writing, 12*(1), 10-25. https://doi.org/10.1016/j.asw.2007.02.002

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6,* 145-178.

Weigle, S. C. (2002). *Assessing writing.* New York, NY: Cambridge University Press.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9,* 27-55.

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly, 37*(2), 345-354. https://doi.org/10.2307/3588510

Williams, H. (2012), "Cohesion and pragmatic theory in second-language writing instruction". *Language and Linguistics Compass, 6*(1), 768-776. doi:10.1111/lnc3.12005

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing, 34,* 16-36. https://doi.org/10.1016/j.asw.2017.08.002

Yang, H. (2014). Toward a model of strategies and summary writing performance. *Language Assessment Quarterly, 11*(4), 403-431. https://doi.org/10.1080/15434303.2014.957381

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education, 23*(1), 31-48.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53-67. https://doi.org/10.1016/j.jslw.2015.02.002

### Author biodata

**Rurik Tywoniw** is a Ph.D. student in Applied Linguistics at Georgia State University. His research interests include second language assessment, second language literacy, and computational linguistics. His work at Georgia State University includes coordinating the Georgia State Test of English Proficiency and teaching Elementary Japanese.

**Scott Crossley** is Professor at Georgia State University. His interests include computational linguistics, corpus linguistics, cognitive science, discourse processing, and discourse analysis. His primary research focuses on the development and application of computational tools in second language learning and text comprehensibility.

## Appendix A

Descriptive statistics for cohesion indices included in the MANOVA between writing tasks.

| | Integrated | | Independent | |
|---|---|---|---|---|
| | M | SD | M | SD |
| Local Cohesion | | | | |
| Amount of repeated use of verb lemmas across two adjacent sentences | 0.031 | 0.016 | 0.032 | 0.013 |
| Number of sentences with repeated use of verb lemmas across two adjacent sentences | 0.545 | 0.265 | 0.544 | 0.230 |
| Number of sentences with repeated use of noun and pronoun lemmas across two adjacent sentences | 0.665 | 0.241 | 0.610 | 0.213 |
| Amount of repeated use of content word lemmas across three adjacent sentences | 0.207 | 0.057 | 0.156 | 0.047 |
| Amount of repeated use of function word lemmas across three adjacent sentences | 0.159 | 0.046 | 0.167 | 0.045 |
| Amount of repeated use of noun and pronoun lemmas across three adjacent sentences | 0.092 | 0.038 | 0.082 | 0.033 |
| Number of sentences with repeated use of adjective lemmas across three adjacent sentences | 0.330 | 0.281 | 0.192 | 0.176 |
| Number of sentences with repeated use of personal pronouns lemmas across three adjacent sentences | 0.191 | 0.196 | 0.433 | 0.240 |
| Use of simple subordinators | 0.044 | 0.017 | 0.037 | 0.015 |
| Use of connectives of addition | 0.025 | 0.012 | 0.028 | 0.012 |
| Total number of additive connectors | 0.044 | 0.015 | 0.047 | 0.014 |
| Use of logical connectives | 0.048 | 0.018 | 0.052 | 0.017 |
| Use of negative logical connectors | 0.010 | 0.008 | 0.007 | 0.006 |
| Total number of connectives | 0.091 | 0.022 | 0.097 | 0.019 |
| **Global Cohesion** | | | | |
| Average amount of repeated verb lemma use per sentence across two adjacent paragraphs | 1.534 | 0.991 | 2.300 | 1.217 |
| Average amount of repeated function word lemma use per sentence across three adjacent paragraphs | 6.593 | 3.214 | 10.937 | 4.236 |
| Average amount of repeated verb lemma use per sentence across three adjacent paragraphs | 0.048 | 0.027 | 0.054 | 0.022 |
| Amount of repeated noun and pronoun lemma use across three adjacent paragraphs | 0.089 | 0.046 | 0.079 | 0.033 |
| Average amount of repeated use of nouns and pronouns per sentence across three adjacent paragraphs | 4.395 | 2.503 | 5.529 | 2.697 |
| **Textual Cohesion** | | | | |
| Lemma type-token ratio | 0.501 | 0.065 | 0.445 | 0.064 |
| Bigram type-token ratio | 0.882 | 0.048 | 0.875 | 0.054 |
| Use of determiners | 0.146 | 0.034 | 0.097 | 0.027 |
| Repeated use of pronouns throughout a text | 0.375 | 0.061 | 0.418 | 0.072 |

## Appendix B

MANOVA findings for cohesion differences between writing tasks.

| | *F* (1, 479) | *p* | Partial $\eta^2$ | *d* |
|---|---|---|---|---|
| **Local Cohesion** | | | | |
| Amount of repeated use of verb lemmas across two adjacent sentences | 2.200 | 0.138 | 0.002 | 0.206 |
| Number of sentences with repeated use of verb lemmas across two adjacent sentences | 0.006 | 0.939 | 0.000 | 0.141 |
| Number of sentences with repeated use of noun and pronoun lemmas across two adjacent sentences* | 13.804 | < 0.001 | 0.014 | 0.123 |
| Amount of repeated use of content word lemmas across three adjacent sentences* | 230.371 | < 0.001 | 0.194 | 0.191 |
| Amount of repeated use of function word lemmas across three adjacent sentences | 8.763 | 0.073 | 0.009 | 0.022 |
| Amount of repeated use of noun and pronoun lemmas across three adjacent sentences* | 19.948 | < 0.001 | 0.020 | 0.140 |
| Number of sentences with repeated use of adjective lemmas across three adjacent sentences* | 83.711 | < 0.001 | 0.080 | 0.448 |
| Number of sentences with repeated use of personal pronouns lemmas across three adjacent sentences* | 292.859 | < 0.001 | 0.234 | 0.201 |
| Use of simple subordinators* | 55.048 | < 0.001 | 0.054 | 0.125 |
| Use of connectives of addition* | 17.151 | < 0.001 | 0.018 | 0 |
| Total number of additive connectors | 7.588 | 0.006 | 0.008 | 0.069 |
| Use of logical connectives* | 15.243 | < 0.001 | 0.008 | 0.057 |
| Use of negative logical connectors* | 40.160 | < 0.001 | 0.040 | 0.283 |
| Total number of connectives* | 16.613 | < 0.001 | 0.017 | 0.146 |
| **Global Cohesion** | | | | |
| Average amount of repeated verb lemma use per sentence across two adjacent paragraphs * | 114.523 | < 0.001 | 0.107 | 0.204 |
| Average amount of repeated function word lemma use per sentence across three adjacent paragraphs* | 320.363 | < 0.001 | 0.251 | 0.272 |
| Average amount of repeated verb lemma use per sentence across three adjacent paragraphs* | 12.129 | < 0.001 | 0.013 | 0.203 |
| Amount of repeated noun and pronoun lemma use across three adjacent paragraphs* | 15.179 | < 0.001 | 0.016 | 0.325 |
| Average amount of repeated use of nouns and pronouns per sentence across three adjacent paragraphs* | 45.574 | < 0.001 | 0.045 | 0.075 |
| **Textual Cohesion** | | | | |
| Lemma type-token ratio* | 183.415 | < 0.001 | 0.161 | 0.016 |
| Bigram type-token ratio | 5.272 | 0.022 | 0.005 | 0.117 |
| Use of determiners* | 635.011 | < 0.001 | 0.399 | 0.228 |
| Repeated use of pronouns throughout a text* | 99.522 | < 0.001 | 0.094 | 0.165 |

*Index was included in the Discriminant Function Analysis

Note: Partial η2 is the effect size derived within the MANOVA model. Cohen's d is the effect size of pairwise comparisons.