Behavior Research Methods (2019) 51:14–27
https://doi.org/10.3758/s13428-018-1142-4

The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap

Scott A. Crossley[1] · Kristopher Kyle[2] · Mihai Dascalu[3]

Published online: 8 October 2018

Abstract

This article introduces the second version of the Tool for the Automatic Analysis of Cohesion (TAACO 2.0). Like its predecessor, TAACO 2.0 is a freely available text analysis tool that works on the Windows, Mac, and Linux operating systems; is housed on a user's hard drive; is easy to use; and allows for batch processing of text files. TAACO 2.0 includes all the original indices reported for TAACO 1.0, but it adds a number of new indices related to local and global cohesion at the semantic level, reported by latent semantic analysis, latent Dirichlet allocation, and word2vec. The tool also includes a source overlap feature, which calculates lexical and semantic overlap between a source and a response text (i.e., cohesion between the two texts based measures of text relatedness). In the first study in this article, we examined the effects that cohesion features, prompt, essay elaboration, and enhanced cohesion had on expert ratings of text coherence, finding that global semantic similarity as reported by word2vec was an important predictor of coherence ratings. A second study was conducted to examine the source and response indices. In this study we examined whether source overlap between the speaking samples found in the TOEFL-iBT integrated speaking tasks and the responses produced by test-takers was predictive of human ratings of speaking proficiency. The results indicated that the percentage of keywords found in both the source and response and the similarity between the source document and the response, as reported by word2vec, were significant predictors of speaking quality. Combined, these findings help validate the new indices reported for TAACO 2.0.

Keywords Cohesion · Coherence · Natural language processing · Essay quality · Speaking proficiency

Cohesion generally refers to text elements that help readers facilitate comprehension (Graesser, McNamara, & Louwerse, 2003). These elements can include argument overlap across sentences and paragraphs, the use of connectives to link sentence and paragraphs, and casual relationships within a text. Cohesion can be categorized at a base level as either local or global, where *local* cohesion refers to textual links between sentences, and *global* cohesion identifies links between larger text segments such as paragraphs (Givón, 1995; Kintsch, 1995). Measuring text cohesion is important because it can help identify text that is more difficult to comprehend, text genres, and writing quality (Crossley & McNamara, 2010, 2011; Crossley, Russell, Kyle, & Römer, 2017a; Gernsbacher, 1990). A few natural language processing (NLP) tools exist that measure cohesion, including Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004), the Tool for the Automatic Analysis of Cohesion (TAACO 1.0; Crossley, Kyle, & McNamara, 2016), and ReaderBench (Dascalu, McNamara, Trausan-Matu, & Allen, 2018).

This study introduces the Tool for the Automatic Analysis of Cohesion (TAACO) 2.0. Like its predecessor (TAACO 1.0), the tool provides hundreds of automatically computed linguistic features related to text cohesion. The tool is available for the Windows, Mac, and Linux operating systems, is housed on a user's hard drive, is easy to use, and allows for batch processing of .txt files. Like the original TAACO, the tool incorporates both classic and recently developed indices. The tool includes all of the original features reported for TAACO 1.0, including local (i.e., sentence-level connections), global (i.e., paragraph-level connections), and overall text cohesion indices based on connective lists, words, lemmas, and synonym overlap, along with type–token indices. TAACO 2.0 adds to the original tool by integrating semantic similarity features based on latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998), latent

✉ Scott A. Crossley
sacrossley@gmail.com

[1] Department of Applied Linguistics/ESL, Georgia State University, Atlanta, GA, USA

[2] Department of Second Language Studies, University of Hawai'i at Manoa, Honolulu, Hawaii, USA

[3] Department of Computer Sciences, Politehnica University of Bucharest, Bucharest, Romania

Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003), and word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), at both the local and global levels. In addition, the tool includes a feature that allows users to examine text overlap in terms of semantic similarity and keywords between a source text and a response text (e.g., a text that integrates a source document and the actual source document).

We assessed the validity of TAACO 2.0 in two studies. The first investigated how semantic overlap features at the local and global levels can assess writing quality, both alone and in conjunction with a number of fixed factors known to be important predictors of writing quality and/or with confounding variables, including prompt, sex, age, and reading and computer habits. In the second study, we examined whether overlap between speaking tasks and student responses was predictive of overall judgments of speaking proficiency, which is partially predicated on discourse that exhibits coherent and efficient expression of relevant ideas. Like the first study, this study also controlled for a number of individual differences and confounding variables attested to be important in speaking quality, including working memory, gender, age, language proficiency, and task differences (e.g., topic). Thus, our goal was not only to introduce TAACO 2.0 but also to validate its ability to model human judgments of language proficiency.

Cohesion

Cohesion references how connected discourse segments are to one another, on the basis of textual features. Cohesion helps individuals interpret parts of a discourse because it indicates dependencies among the text elements (Halliday & Hasan, 1976). Discourse cohesion is an important feature of discourse comprehension, especially for more challenging discourse with increased knowledge demands (Loxterman, Beck, & McKeown, 1994; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996). Traditionally, cohesion has been studied in terms of overt connections between text segments. For instance, Halliday and Hasan identified a number of devices that connect pieces of text together, most of which are explicitly located in a text. These cohesive devices included word substitution, reference (generally through pronouns), use of conjunctions, and lexical repetition. The latter phenomenon, also referred to as *lexical cohesion*, can be both explicit and implicit (Sanders & Maat, 2006). For instance, if two adjacent sentences contain the same noun (e.g., *dog*), the lexical repetition will help link the sentences explicitly. However, if the first sentence contains the word *dog* and the second sentence contains the word *fur*, the explicitness of the connection weakens considerably. Sometimes the links between two sentences weaken to the point that the systematic relationship between shared lexical items is difficult to evaluate. At that point, connectedness in terms of cohesion features breaks, which may affect the mental representation of the discourse that the reader develops (McNamara et al., 1996; O'Reilly & McNamara, 2007; Sanders & Maat, 2006). Cohesion can also be considered in larger discourse structures that cross document boundaries. For instance, research examining source-based (i.e., integrated) writing and speaking assessments may consider cohesive properties between a student's response and the source documents that the student is expected to integrate into the response. Stronger links between a student's response and the source documents should increase interdocument cohesion, and likely will help expert raters and readers develop a more coherent representation of a student's response (Crossley, Clevinger, & Kim, 2014).

The mental representation that develops as a result of cohesion in the discourse is referred to as *coherence* (McNamara et al., 1996; O'Reilly & McNamara, 2007; Sanders & Maat, 2006). The distinction between cohesion and coherence is crucial to understanding the importance of cohesion devices. Whereas *cohesion* refers to the presence or absence of explicit cues in the text that allow the reader to make connections between the ideas in the text, *coherence* refers to the understanding that each individual reader or listener derives from the discourse (i.e., the coherence of the text in the mind of the reader). Whereas cohesion has a strongly linguistic function and can be measured on the basis of the linguistic features present in the text, coherence may rely on linguistic features, but it may also strongly interact with language proficiency and background knowledge (McNamara et al., 1996; O'Reilly & McNamara, 2007). Thus, coherence is not necessarily a property of textual features. For instance, local cohesion assists readers with low background knowledge to understand texts (McNamara et al., 1996), but not more experienced readers, who may benefit from less cohesion that forces them to bridge cohesion gaps in the text, using such strategies as inferencing. Cohesion features also interact differently in terms of both register and interlocutors. For instance, grade level affects the production of cohesive devices (Crowhurst, 1987; Fitzgerald & Spiegel, 1983; Yde & Spoelders, 1985), with the distance between the referents being used to create cohesive links decreasing as a function of grade (Fitzgerald & Spiegel, 1983; McCutchen & Perfetti, 1982; Yde & Spoelders, 1985). In addition, there is a movement away from the production of cohesive devices and a preference for more complex syntactic structures with increasing grade level (Haswell, 2000; McCutchen & Perfetti, 1982).

Assessing discourse cohesion

There are a number of ways to assess discourse cohesion and its effects on comprehension and processing. Perhaps the most common approach is to assess links between human ratings of discourse cohesion and/or quality. As an example, Crossley

and McNamara (2010, 2011) examined the associations between local cohesion features calculated by the computational tool Coh-Metrix (Graesser et al., 2004) and human judgments of text coherence in a corpus of independent essays written by freshman college students. In the first study, Crossley and McNamara (2010) found strong correlations between the human judgments of writing quality and coherence, even while the indices that calculated local cohesion features (e.g., connectives, anaphoric reference, and lexical overlap) demonstrated negative correlations with ratings of text coherence. In the second study, Crossley and McNamara (2011) used both local and global features of cohesion to assess ratings of text coherence. These features were reported by the Writing Assessment Tool (WAT; Crossley, Roscoe, & McNamara, 2013) and, as in the first study, negative correlations were reported between local cohesion devices and human judgments of coherence. However, positive correlations were reported between global indices of cohesion, which calculated semantic overlap between the initial, middle, and final paragraphs of the essays, and judgments of text coherence. Similar trends have been reported for studies examining links between local and global cohesion devices and human judgments of writing quality (as compared to judgments of coherence). For instance, a number of studies have found that local cohesion devices are not strong predictors of essay quality (Evola, Mamer, & Lentz, 1980; McCulley, 1985; Neuner, 1987; although see Faigley & Witte, 1981, for counter evidence), and at least one study has shown strong links between global cohesion devices and essay quality (Neuner, 1987).

These findings, however, may not hold for younger students, who often depend on local cohesion devices to create coherent texts. These students use a greater number of explicit cohesion features, such as referential pronouns and connectives (King & Rentel, 1979). With time, developing writers begin to produce fewer explicit cohesion cues as a strategy to organize their texts (McCutchen, 1986; McCutchen & Perfetti, 1982). Moreover, the trends reported above begin to hold in writing at this level (i.e., the production of explicit local cohesion cues is commonly associated with less proficient writing; Crossley, Weston, Sullivan, & McNamara, 2011; McNamara, Crossley, & McCarthy, 2010; McNamara, Crossley, & Roscoe, 2013). The findings that local cohesion features are not positively associated with writing quality may also not hold for second language (L2) writers. A number of studies have reported positive associations between essay quality in L2 writing and the use of local cohesive devices, including cohesive ties (Jafarpur, 1991), sentence transitions (Chiang, 2003), and other text cohesion devices (Liu & Braine, 2005). However, more recent studies using advanced NLP tools have reported the opposite effect (i.e., that local cohesion devices are negatively associated with L2 essay quality). For instance, Crossley and McNamara (2012) reported that local cohesion devices (i.e., lexical overlap between sentences and the use of positive logical connectives) demonstrated negative correlations with ratings of writing quality. Similarly, Guo, Crossley, and McNamara (2013) found that the production of local cohesion devices (e.g., content word overlap, and conditional connectives) was negatively correlated with essay quality ratings. In total, these studies have tended to indicate that local cohesion devices are not related or negatively related to human judgments of coherence and writing quality, whereas global cohesion features are positively related to ratings of coherence and writing quality (at least for adult writers).

A number of studies have recently begun to examine the relationships between cohesion devices and speaking ability, as well. Much of this work has aligned with L2 acquisition studies that have focused on the development of L2 speaking proficiency for both integrated (i.e., source-based speaking) and independent tasks. In an early study, Crossley and McNamara (2013) examined the links between lexical and cohesive features in independent speaking responses and human ratings of quality. They reported strong and positive associations between causal cohesion and human ratings of quality. They also examined cohesion between the prompt and the response (i.e., keyword overlap between the two), and found that speakers who used more keyword types from the prompt were scored higher. In other studies, Crossley, Clevinger, and Kim (2014) and Crossley and Kim (in press) investigated cohesive links between source texts and spoken responses in integrated writing tasks. They found that words integrated from the source text into the speaking response were more likely to be repeated in the source text and more likely to be found in phrases with positive connectives. In addition, words integrated from the source texts into the responses were highly predictive of speaking quality. Finally, Kyle, Crossley, and McNamara (2016) examined differences between L2 speakers' responses to both independent and integrated speaking tasks and found that independent speaking tasks result in less given information (i.e., information available previously in the discourse). Although studies examining cohesion and spoken discourse in terms of comprehension and processing are infrequent, the existing studies above indicate that cohesion features in discourse, as well as cohesive links between the speaking prompt/source text and a response, are important components of speaking quality.

TAACO 2.0

TAACO 1.0 was a user-friendly tool written in Python that allowed users with limited computer programming knowledge to examine the cohesion features of a text using an intuitive graphical user interface (GUI). The tool incorporated a part-of-speech (POS) tagger from the Natural Language Tool Kit (Bird, Klein, & Loper, 2009) and synonym sets from the WordNet lexical database (Miller, 1995). At the time,

TAACO 1.0 differed from other automatic tools that assessed cohesion because it reported on a variety of local and global features of cohesion, including connectives, text givenness, type–token ratios (TTRs), lexical overlap at the sentence and paragraph levels, and synonym overlap at the local and global levels. Moreover, TAACO 1.0 was housed on the user's hard drive and allowed for batch-processing of texts.

The major additions included in TAACO 2.0 comprise the integration of semantic similarity features calculated at the local and global levels, as well as an overlap measure between two given texts using both semantic similarity and keyword features. Whereas local and global cohesion indices have commonly been examined in cohesion studies, analyses of similarities between prompts/source texts and response texts are less common. In the process of further developing TAACO, a number of other small changes were made to the manner in which features were counted. These included refined distinctions between content words (nouns, lexical verbs, adjectives, and adverbs that are derived from adjectives), function words (all other words), and the use of a dependency parser (Chen & Manning, 2014) to disambiguate word forms that can be used as both cohesive devices and for other purposes. For example, the word "*so*" can be used as a connective (e.g., "They were out of pizza, *so* we ate sushi instead.") or as an intensifier ("That sushi was *so* good"), which can be disambiguated using the syntactic dependency parser. Below we discuss the major additions found in TAACO 2.0.

Semantic similarity features

An important element of discourse cohesion in terms of NLP techniques is to what extent computational models of semantic memory (Cree & Armstrong, 2012) are capable of highlighting underlying semantic relations in texts. These computational models rely on unsupervised learning methods that measure cohesion between textual fragments (Bestgen & Vincze, 2012). Common models include semantic vector spaces using LSA (Landauer et al., 1998), topic distributions in LDA (Blei et al., 2003), and word2vec vector space representations (Mikolov et al., 2013).

Latent semantic analysis LSA (Landauer et al., 1998) is a mathematical optimization for representing word meanings in an orthogonal vector space created through an unsupervised learning method applied on a large text corpus. The vector space model is based on word co-occurrences within documents that establish relationships between concepts. LSA builds a term–document matrix, which is normalized using log-entropy, followed by singular-value decomposition (SVD; Golub & Reinsch, 1970) and projection over the most representative k dimensions. LSA has been subjected to extensive psychological experiments (Landauer, McNamara, Dennis, & Kintsch, 2007), which have suggested that it successfully models human linguistic and cognitive knowledge. The LSA model used in TAACO 2.0 was created using the stochastic SVD from Apache Mahout (Owen, Anil, Dunning, & Friedman, 2011), and it considered several optimizations, including ignoring stop words and nondictionary words, reducing inflected word forms to their corresponding lemmas, erasing words with low frequencies (i.e., five occurrences), elimination of paragraphs with fewer than 20 words, and projection over 300 dimensions, as suggested by Landauer et al. (2007).

Latent Dirichlet allocation LDA (Blei et al., 2003) is a generative probabilistic process in which documents are perceived as mixtures of multiple topics. Documents and words alike are topic distributions drawn from Dirichlet distributions, and topics (i.e., latent variables in the model) are perceived as Dirichlet distributions over the input vocabulary (Kotz, Balakrishnan, & Johnson, 2004). Similar to LSA, related concepts have similar topic probabilities based on underlying co-occurrence patterns. Our LDA model was trained using Mallet (McCallum, 2002) over 300 topics (to ensure comparability with the other models).

Word2vec Word2vec (Mikolov et al., 2013) relies on a neural-network model to represent words and phrases in a vector-space model with a limited number of dimensions, k. Each word's embedding is computed using the context around it within the training dataset; thus, words co-occurring in similar contexts are represented as being closer, whereas words with dissimilar contexts are represented as being farther apart, in different regions of the vector space. Our word2vec model was trained using GenSim (Řehůřek & Sojka, 2010) with Skip-Gram negative sampling, a window of five words, and 300 dimensions. It should be noted that word2vec is based on a matrix factorization similar to the SVD used in developing the LSA spaces, which might lead to multicollinearity among indices derived from the two spaces (Levy & Goldberg, 2014).

Semantic similarity calculations in TAACO 2.0 For each of these models (LSA, LDA, word2vec), TAACO 2.0 calculates the average similarity between progressive adjacent segments (sentences or paragraphs) in a text. All semantic models in TAACO 2.0 were trained on the newspaper and popular magazine sections of the Corpus of Contemporary American English (COCA; Davies, 2008). The magazine section includes articles related to news, health, home, religion, and sports (among others). The newspaper section includes articles taken from ten newspapers in the United States and includes local news, opinion, and sports. The total size of these two corpora is around 219 million words (~ 110 million words in the magazine corpus and ~ 109 million words in the newspaper corpus). Prior to developing the semantic spaces, all function words were removed, as were all identified non-English words (e.g., misspellings). Finally, all words in the corpora were lemmatized.

All semantic models relied on the bag-of-words assumption, in which word order is disregarded. For LSA and word2vec, similarity scores were determined using the cosine similarity between the vectors corresponding to the compared segments (e.g., sentences, paragraphs, or documents), which were obtained from summing the vector weights for each word in a particular segment. Average vector weights were calculated by considering the square root of the sum of the squares of vector weights. The final cosine similarity score consisted of the sum of the products of the summed vector weights from both segments, divided by the product of the average vector weights from both segments. LDA relatedness scores were calculated as the inverse of the Jensen–Shannon divergence between the normalized summed vector weights for the words in each segment.

TAACO semantic similarity indices are calculated using two methods for exploring adjacent text segments (i.e., sentences and paragraphs). In the first, similarity scores are calculated between Segments (e.g., sentences) 1 and 2, 2 and 3, 3 and 4, and so forth, until all progressive segment pairs have been examined. In the second method, which accounts for similarity beyond a single adjacent segment, similarity scores are calculated between Segment (e.g., sentence) 1 and the combination of Segments 2 and 3, then Segment 2 and the combination of Segments 3 and 4, and so forth, until all progressive segments have been compared. Additionally, for each semantic similarity calculation method outlined above, a single source similarity index is calculated between a source text and a target text (e.g., an essay that integrates information from a source document and the source document itself).

Keyword overlap features

Source similarity indices provide an overall evaluation of the similarity between the words in a source text and a target text, but they do not differentiate between words that are paraphrased, words that are directly copied, and words that are text-prominent. Keyword overlap indices measure the degree to which important words and *n*-grams (i.e., the bigrams, trigrams, and quadgrams of *n* consecutive words) from the source text are found in the target text. TAACO 2.0 identifies keywords and *n*-grams by comparing the relative frequency of these items in the source text to the relative frequency of the same items in the news and magazine sections of COCA. Any word or *n*-gram that occurs at least twice in the source text and that occurs more frequently in the source text than in the reference corpus (using normed frequency counts) is preliminarily selected as a keyword or *n*-gram. After the preliminary list has been compiled, the top 10% of words or *n*-grams (i.e., those that occur with the highest frequency in the source text as compared to COCA) are selected as keywords or *n*-grams. TAACO 2.0 then calculates the proportion of words or *n*-grams in each target text that are keywords or *n*-grams in the source text. In total, TAACO 2.0 calculates 23 keyword overlap indices, as described below.

TAACO 2.0 calculates keyword overlap for all words, bigrams, trigrams, and quadgrams. Part-of-speech (POS) statistics, considering sensitive key overlap indices at the word level for nouns, adjectives, and verbs, are also computed. Also included are key overlap indices for words that are verbs, nouns, adjectives, or adverbs. At the *n*-gram level, POS sensitive keyword overlap indices are calculated that allow for variable slots within the *n*-gram. For these indices, only words with particular parts of speech (e.g., nouns) are realized as concrete items, whereas other words are counted as variable slots. For example, the bigram "generalized reciprocity," which consists of an adjective ("generalized") and a noun ("reciprocity"), would be represented as "X reciprocity" (wherein X represents a variable slot). Such indices are available with the following parts of speech, included as concrete items: nouns, adjectives, verbs, verbs and nouns, as well as adjectives and nouns.

Study 1

Method

The data used in this study were originally reported in Crossley and McNamara (2016). In that study we used analyses of variance (ANOVAs) to examine the effects that prompt, essay elaboration, and enhanced cohesion had on expert ratings of essay quality. Two essay prompts similar to those used in the SAT were selected after consulting expert teachers. The prompts concerned fitting into society and the importance of winning.

Study design Student participants were recruited from Introductory Psychology courses at a Midwestern university in the United States (N = 35). During data collection, demographic and survey data were collected from the students. The students included 27 females and eight males. The reported ethnic makeup for the students was 31.5% African American, 40.5% Caucasian, and 5.5% who identified themselves as "Other." The students' average age was 19 years, with an age range of 17 to 31. The survey collected information about the number of writing courses completed, hours on the Internet, hours spent sending e-mails, hours spent using word-processing software, and how much the students enjoyed reading and writing. The students were then informed that they would write two original essays on the two different prompts using a laptop computer.

The first essay assignment and corresponding prompt were presented at the top of an open text document on the computer and were not visible to students until all instructions had been given. Students were told to develop and write their response to the first prompt using existing knowledge. They were not allowed to use notes or the Internet, or to ask questions about what to write. Each student was allotted 25 min to compose

the *original* response to the prompt. The student was then provided an additional 15 min to elaborate on the essay just written. Students were asked to add at least two additional paragraphs of about four or five sentences that would provide additional examples to support the main idea in the essay. The second essay prompt was then presented to the student, and the same procedure was followed. The order of essay prompts (Fitting in; Winning) was counterbalanced across students. Students were not told beforehand that they would be given extra time to revise the essays they had written.

Cohesion revisions An expert in discourse comprehension revised each of the four essays composed by the students (the two original essays and the two elaborated essays) to increase text cohesion within the essay at both the local and global levels. Text cohesion was increased using two methods. First, cohesion within the essay was increased by adding word overlap across sentences and paragraphs (e.g., linking text segments together by the repetition of words). Increasing the links between sentences was meant to increase local cohesion, while increasing links between the paragraphs was meant to increase global cohesion. Second, if unattended demonstratives were used (e.g., "*this* is good" or "*that* demonstrates the point"), the referents to those anaphors (the *this* and *that*) were specified. No other modifications were made to the essays. All changes were made such that the writer's original meaning, mechanics, voice, and word choices were maintained. The modifications made were checked by a second expert for accuracy, and any disagreements were remediated between the two raters until agreement was reached.

Corpus The essays collected during the experiment formed the corpus used in this study. The full corpus comprised 280 essays written by 35 freshman students. The essays included the two original essays on two different essay prompts written by the students as well as the revised version of each essay created by the students to elaborate on the original content. As we noted above, these four essays were each revised by a discourse expert in order to increase the cohesion of the essay, resulting in eight essays per student in the corpus: two original essays, two original essays with elaboration (by the student), two original essays with improved cohesion (by an expert), and two essays with elaboration (by the student) and improved cohesion (by an expert).

Essay evaluation The essays were then evaluated by expert composition raters. Twelve raters with at least one year's experience teaching composition classes at a large university rated the 280 essays in the corpus using a holistic grading scale based on a standardized rubric commonly used in assessing SAT essays (see the Appendix for the SAT rubric) and a rubric that assessed individual features of the text, including a text coherence feature. The text coherence feature was labeled *Continuity* and was defined as "The essay exhibits coherence throughout the essay by connecting ideas and themes within and between paragraphs." The holistic grading scale and the rubric scales each had a minimum score of 1 and a maximum score of 6. The raters were informed that the distances between scores were equal. Accordingly, a score of 5 was as far above a score of 4 as a score of 2 was above 1.

The raters worked in pairs and had undergone previous training. Each rating pair evaluated a selection of 70 essays from the corpus. The essays were counterbalanced, such that each group did not score more than one essay from each writer on each prompt and such that each group scored a similar number of essays from each essay type (original, original with elaboration, original with added cohesion, and original with elaboration and cohesion). The raters were blind to condition as well as to the variables of focus in the study. Interrater reliability was acceptable between the raters ($r = .73$).

TAACO 2.0 indices To assess the modifications made to the essays, we used TAACO 2.0. We were specifically interested in the new semantic similarity indices provided by the LSA, LDA, and word2vec semantic spaces available in TAACO. For this analysis, we selected LSA, LDA, and word2vec indices for both sentence and paragraph overlap. We selected indices that calculated overlap across two elements (i.e., across two sentences or paragraphs).

Statistical analysis We used linear mixed effects (LME) models in R (R Core Team, 2017) using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) to examine the relationships between the semantic similarity measures reported in TAACO 2.0 and human ratings of text cohesion. Unlike the original study, which had used repeated measure ANOVAs, the LME models used here accounted for both pooled and individual variance among participants, as opposed to one pooled group, by including subjects as random effects (i.e., assigning a unique intercept for each participant).

Prior to running an LME model, we examined the correlations between the semantic similarity indices and the human scores for cohesion in order to select variables for inclusion in the models that reported at least a small effect size ($r > .100$) and that were not multicollinear. We set multicollinearity at $r > .700$. We conducted two LME models. The first one examined the effects of the new semantic similarity variables that were not multicollinear on the human ratings of cohesion and included subjects as random effects. The second LME model was conducted to examine the effects of these cohesion features when controlling for other fixed factors related to writing quality. For this model, we entered the human rating of cohesion as the dependent variable and entered the semantic similarity variables that were not multicollinear, the condition (original essay, original essay with elaboration, original essay with added cohesion, and original essay with elaboration and added cohesion), the prompt, order of writing (first or second essay), sex, ethnicity,

age, and the survey results for reading and computer use as fixed factors. As in the first analysis, we included subjects as random effects. We used `lmerTest` (Kuznetsova, Brockhoff, & Christensen, 2016) to derive p values from the models, `multcomp` (Hothorn, Bretz, & Westfall, 2008) to obtain full pairwise comparisons between the four conditions, and the `MuMIn` package (Nakagawa & Schielzeth, 2013) to obtain marginal and conditional R^2 measuring.

Results

Pearson correlations indicated that all paragraph-level variables were correlated at above $r > .700$. The strongest correlation with essay scores was for word2vec, and as a result, this variable was selected for inclusion in the LME model. None of the sentence-level semantic similarity indices were multicollinear, and all showed correlations with the response variable that were $r > .100$, so all sentence-level variables were included in the LME model. A correlation matrix for all selected variables is reported in Table 1.

An LME model using only the linguistic features reported a significant effect only for word2vec paragraph similarity. No significant effects were reported for any of the sentence-level semantic similarity indices. A second LME model was conducted using only the significant fixed effect of word2vec paragraph semantic similarity. This model reported a marginal R^2 of .148 and a conditional R^2 of .470. Table 2 reports the intercept and fixed effect, including their coefficients, standard errors, t values, and p values.

A full LME model using speaking proficiency scores as the dependent variable reported significant effects for condition, essay order, and word2vec paragraph similarity. No significant effects were reported for prompt, sex, ethnicity, age, or the survey results for reading and computer use. A second LME model was conducted using only the significant fixed effects of condition, order, and word2vec paragraph semantic similarity. The model reported a marginal R^2 of .176 and a conditional R^2 of .510. Table 3 reports the intercept and fixed effect, including their coefficients, standard errors, t values, and p values. The results indicated that second essays were scored higher than first essays and that the original essays with modified cohesion were scored higher for continuity than both the original and elaborated essays. Finally, essays that had greater global cohesion, as found in semantic similarity (word2vec) across paragraphs, were scored higher.

Table 2 LME results for prediction of continuity score (only cohesion features) in Study 1

Fixed Effect (baseline)	Estimate	Std. Error	t	p
Intercept	3.726	0.075	49.696	< .001
word2vec paragraph	0.274	0.036	7.589	< .001

Discussion

Previous research has demonstrated that global cohesion features are important predictors of essay quality and text coherence. For instance, Crossley and McNamara (2011) reported modest correlations between semantic similarity indices measured between paragraphs using LSA and human ratings of text quality. Their study used WAT, an NLP tool currently under development and not yet available for public use. The results of the present study support these findings, in that semantic overlap at the paragraph level, as measured by LSA, word2vec, and LDA, showed strong correlations with ratings of text coherence. However, all these indices were extremely multicollinear with one another, and only a single index could be used for our analyses (word2vec). This multicollinearity likely existed because differences between the models were reduced across larger segments of text as a result of most paragraphs having semantic links with each other. When used in an LME model, word2vec paragraph overlap explained almost 15% of the variance in human ratings of text coherence, indicating that essays with greater semantic overlap across paragraphs were rated as being more coherent. Word2vec was also a significant predictor in a full model when a number of other fixed effects were considered, including demographics, survey results, condition, and essay order. This second model indicated that semantic similarity across paragraphs was a positive predictor of judgments of essay coherence. In addition, the initial essays were scored lower than the subsequent essays, and the essays modified to increase cohesion were scored higher on text coherence than were the other essay types.

Of note, no sentence-level cohesion variables were included in the LME models, indicating that they explained no unique

Table 1 Pearson correlations: Continuity scores and semantic similarity indices in Study 1

Variable	Continuity Score	LSA Paragraph	LDA Paragraph	word2vec Paragraph	LSA Sentence	LDA Sentence
LSA paragraph	.464					
LDA paragraph	.459	.958				
word2vec paragraph	.484	.977	.989			
LSA sentence	.185	.196	.128	.143		
LDA sentence	.200	.134	.117	.138	.191	
word2vec sentence	.348	.373	.327	.357	.456	.676

Table 3 LME results for prediction continuity score (all fixed factors) in Study 1

Fixed Effect (baseline)	Estimate	Std. Error	*t*	*p*
Intercept	2.840	0.138	20.617	< .001
Elaborated condition (baseline original)	0.237	0.087	2.733	< .010
Cohesion condition (baseline original)	0.209	0.083	2.499	< .050
Cohesion and elaborated condition (baseline original)	0.304	0.090	3.376	< .001
Cohesion condition (baseline elaborated)	0.221	0.085	2.592	< .010
Cohesion and elaborated condition (baseline elaborated)	0.126	0.085	1.489	> .050
Cohesion and elaborated condition (baseline cohesion)	– 0.095	0.087	– 1.092	> .050
Essay order	– 0.149	0.057	– 2.604	< .010
word2vec paragraph similarity	0.242	0.038	6.337	< .001

variance beyond the paragraph-level variables. However, unlike for the paragraph-level variables, strong correlations were not reported among the sentence-level variables (see Table 1). The strongest correlation between the sentence-level semantic similarity measures was reported for LDA and word2vec. These correlational analyses indicated greater variation among the types and strengths of semantic links found between sentences (when compared to paragraphs), which makes intuitive sense. Larger segments of texts, by their nature, will increase the likelihood of similarities being reported for semantics, word embeddings, and topics; smaller text segments will decrease this likelihood. Thus, different semantic similarity features calculated at the sentence level should explain unique variance when compared to each other.

In addition, all the correlations reported for the sentence-level variables and the human judgments of text coherence were positive and significant, which is different from the results of previous analyses (Crossley & McNamara, 2010, 2011). In at least one case, a reported correlation showed a medium effect size (i.e., the word2vec index). The likely reason for these differences was the inclusion of LDA and word2vec models within TAACO (as compared to relying on LSA models alone), as well as more principled development of the semantic spaces reported in TAACO. For instance, the LSA space in TAACO uses a much larger and more representative corpus than the corpus used in developing the Coh-Metrix LSA space. The TAACO LSA space (like all TAACO semantic spaces) is based on the newspaper and magazine subcorpora in COCA, which are both larger than the corpus used to design the Coh-Metrix LSA space (~ 219 million words/~ 97 million words after preprocessing in COCA, as compared to ~ 13 million words/~ 7 million words after preprocessing in the Touchstone Applied Sciences Association [TASA] corpus used for Coh-Metrix). The corpora used in the TAACO semantic spaces are also more representative of the language to which adults are exposed than is the TASA corpus, which includes texts appropriate for 3rd, 6th, 9th, and 12th grades and for college-level readers. In addition, the TAACO semantic spaces reported here had stop words and nondictionary words removed, and all words lemmatized. This is in contrast to the Coh-Metrix indices used in previous studies, in which only stop words were removed (see Crossley, Dascalu, & McNamara, 2017b, for a fuller discussion of this topic).

To examine the presumed differences between the TAACO and Coh-Metrix semantic spaces, simple correlations were measured among the continuity score, the TAACO LSA indices (sentence-level and paragraph-level), and the LSA indices reported by Coh-Metrix (at both the sentence and paragraph levels). As is reported in Table 4, stronger correlations were reported for the LSA indices calculated in TAACO than in Coh-Metrix (for both the sentence- and paragraph-level features). Surprisingly, no correlations were reported between the LSA paragraph index calculated by TAACO and either the sentence-level or the paragraph-level LSA indices in Coh-Metrix. Stronger correlations were reported between the LSA sentence index calculated by TAACO and both the sentence-level and paragraph-level LSA indices in Coh-Metrix ($r \sim .60$). Finally, the LSA sentence-to-sentence index and the LSA paragraph-to-paragraph index reported by Coh-Metrix were strongly correlated ($r = .860$), whereas the LSA sentence-to-sentence index and the LSA paragraph-to-paragraph index reported by TAACO were not ($r = .196$). These findings indicate that the LSA-based local and global cohesion features calculated by TAACO measure different constructs. In contrast, the strong correlations between the sentence- and paragraph-level LSA spaces reported by Coh-Metrix indicate that they measure similar constructs.

These correlations indicate that the semantic spaces in TAACO may be better representations of both language exposure and linguistic processing (i.e., lemmatized content words), in that they explain stronger effect sizes with human ratings of coherence than do those reported by Coh-Metrix. Although more evidence is needed, we presume that these differences resulted from the more principled development of the TAACO semantic spaces. In addition, these correlations indicate that the local and global features of cohesion measured by TAACO appear to measure different constructs. This may not be true for the Coh-Metrix local and global indices, which are highly multicollinear.

Table 4 Correlations between LSA indices reported by Coh-Metrix and TAACO

Feature	LSA sentence-to-sentence similarity (Coh-Metrix)	LSA paragraph-to-paragraph similarity (Coh-Metrix)	LSA sentence-to-sentence similarity (TAACO)	LSA paragraph-to-paragraph similarity (TAACO)
Continuity score	.172	.147	.185	.464
LSA sentence-to-sentence similarity (Coh-Metrix)		.86	.603	.075
LSA paragraph-to-paragraph similarity (Coh-Metrix)			.607	.055
LSA sentence-to-sentence similarity (TAACO)				.196

Study 2

Method

In the second study, we examined whether greater cohesion between source documents and speaking responses, as found in the TOEFL-iBT integrated speaking tasks, are predictive of human ratings of speaking proficiency. The basic premise of this analysis was to examine whether increased cohesion between texts positively affects expert raters, possibly increasing rater coherence. Thus, this analysis is not focused on local or global cohesion, but rather on intertextual cohesion. As in our previous analysis, we controlled for a number of individual differences (e.g., working memory, gender, age, and language proficiency) and task differences (e.g., topic).

Integrated speaking tasks The TOEFL-iBT listen/speak integrated tasks ask test-takers first to listen to a spoken source text, such as an academic lecture or a conversation in an academic context. The test-taker then provides a spoken response to a question based on a listening prompt, and the answer is recorded for later assessment. Expert raters then score these speech samples using a standardized rubric that assesses delivery, language use, and topic development. The tasks form part of the overall TOEFL score that participants receive, which is considered a measure of academic English knowledge.

Participants The study included 263 participants who were enrolled in an intensive English program. The participants were recruited from intermediate and advanced English classes to ensure that they had appropriate language skills to take the integrated listen/speak section of the TOEFL-iBT. The participants spoke a number of different first languages. Gender was split about evenly (47% of the participants were male, and 53% were female), and the average age of the participants was 24 years.

Materials The materials included a background survey that collected information including age, gender, and highest educational degree. Participants also took two working memory tests: an aural running span test and a listening span test. The aural running span test was used to provide a working memory test that is not overly dependent on L2 proficiency (Broadway & Engle, 2010). The listening span test was similar to that used in previous L2 acquisition studies (Mackey, Adams, Stafford, & Winke, 2010). Participants also completed an institutional TOEFL exam that included three sections: Listening Comprehension, Structure and Written Expression, and Reading Comprehension. Finally, the participants completed two integrated listen/speak TOEFL iBT speaking tasks from one form of the TOEFL iBT (two forms were used). For each question, students were given 20 s to prepare for their response and 60 s to respond to the prompt. The two forms included four topics: note taking, swimming, fungus, and reciprocity.

Procedure All participants attended two data collection sessions. They completed the institutional TOEFL on Day 1, and then completed the background survey, the two working memory tests, and the two integrated listen/speak tasks from the TOEFL iBT speaking test (listening to a conversation vs. listening to a lecture) on Day 2. On average, the participants spent approximately 2 h in the lab on the first day and 1 h 20 min in the lab on the second day. The order of data collection for the two speaking tasks on Day 2 was counterbalanced and randomly assigned to participants.

Transcription Each spoken response was transcribed by a trained transcriber. The transcriber ignored filler words (e.g., "umm," "ahh") but did include other disfluency features, such as word repetition and repairs. Periods were inserted at the end of each idea unit. All transcriptions were independently checked for accuracy by a second trained transcriber. Any disagreements were remediated until agreement was reached.

Human ratings Two expert TOEFL raters scored each speaking response. The raters used the TOEFL-iBT integrated speaking task rubric, which provides a holistic speaking score based on a 0–4 scale, with a score of 4 being the highest. The rating scale is based on three criteria: language use (i.e., grammar and vocabulary), delivery (i.e., pronunciation and prosody), and topic development (i.e., content and coherence).

TAACO 2.0 indices For this analysis, we used the text overlap and similarity indices reported by TAACO 2.0. The overlap indices compute key *n*-grams in a source text and then calculate the percentage of these key *n*-grams in a response text. The similarity indices measure semantic similarity between a source text and a response text using LSA, LDA, and word2vec.

Statistical analysis As in Study 1, we used LME models in R (R Core Team, 2017) using the `lme4` package (Bates et al., 2015). Prior to running an LME model, we examined the correlations between the semantic similarity and key overlap indices and the human scores for speaking proficiency, to select variables for inclusion in the models. We selected all variables that reported at least a small effect size ($r > .100$) and were not multicollinear ($r > .700$). We set multicollinearity at $r > .700$. We conducted two LME models. The first one examined the effects of the semantic similarity variables and the source overlap indices that were not multicollinear on the human ratings of speaking proficiency; it included subjects as random effects. The second model was conducted to examine the effects of these cohesion features when controlling for other individual and task differences related to speaking proficiency. For this LME model, we entered the human rating of speaking proficiency as the dependent variable and entered topic, text order, age, gender, education level, working memory tests, Institutional TOEFL results, and the source overlap and similarity variables as fixed factors. We did not enter condition (conversation or lecture) or order as fixed effects, because they replicated with topic and would have led to rank-deficient models. As in the first LME model, we included subjects as random effects. We used `lmerTest` (Kuznetsova et al., 2016) to derive *p* values from the models, `multcomp` (Hothorn et al., 2008) to obtain full pairwise comparisons between the four topic conditions, and the `MuMIn` package (Nakagawa & Schielzeth, 2013) to obtain marginal and conditional R^2 measuring.

Results

Pearson correlations indicated that all semantic similarity source overlap variables (LSA, LDA, and word2vec) reported at least a weak effect size ($r > .100$) with speaking score. For the keyword overlap indices, four quadgram indices reported at least a weak effect size ($r > .100$) with speaking score. All semantic similarity indices were highly correlated with each other ($r > .900$). As a result, only the word2vec variable, which reported the strongest correlation with speaking score ($r = .369$), was retained. None of the quadgram indices were multicollinear. A correlation matrix for the four variables used in the LME analysis is reported in Table 5.

An LME model using only the linguistic features reported a significant effect only for word2vec source similarity. No significant effects were reported for any of the other source similarity indices. A second LME model was conducted using only the significant fixed effect of word2vec source semantic similarity. The model reported a marginal R^2 of .016 and a conditional R^2 of .655. Table 6 reports the intercept and fixed effect, including their coefficients, standard errors, *t* values, and *p* values.

A full LME model using the human ratings of cohesion as the dependent variable reported significant effects for topic, TOEFL reading and listening scores, key quadgrams that contained a noun and/or verb, and word2vec source overlap similarity. No significant effects were reported for working memory scores (listening or running span), TOEFL reading score, or the other key quadgram indices. A second LME model was conducted using only the significant fixed effects of topic, TOEFL listening and structure scores, key quadgrams, and word2vec similarity. The model reported a marginal R^2 of .401 and a conditional R^2 of .710. Table 7 reports the intercept and fixed effect, including their coefficients, standard errors, *t* values, and *p* values. The results indicate that summarizing the conversation related to notes led to higher speaking scores than did the other topics, and that the lecture related to fungus led to the lowest speaking score. The results also indicated that participants with higher TOEFL listening and structure scores were scored higher on overall speaking proficiency. In terms of overlap with the source text, participants who produced a greater number of key quadgrams in their responses and who had greater similarity between the source text and the response received higher speaking scores.

Discussion

A number of studies have recently begun to examine the links between cohesive features in speech and speaking proficiency. Most of this work has focused on L2 language acquisition studies

Table 5 Pearson correlations: TOEFL speaking scores and source overlap indices in Study 2

Variable	Speaking score	Source similarity word2vec	Percentage key quadgrams that include a verb	Percentage key quadgrams that include a verb and/or a noun
Source similarity word2vec	.369			
Percentage key quadgrams that include a verb	.150	.215		
Percentage key quadgrams that include a verb and/or a noun	.139	.437	.166	
Percentage key quadgrams that include verb and/or an adjective	.128	.369	.179	.456

Table 6 LME results for predicting TOEFL speaking score (cohesion features only) in Study 2

Fixed Effect (baseline)	Estimate	Std. Error	*t*	*p*
(Intercept)	1.700	0.126	13.443	< .001
Source similarity word2vec	0.958	0.156	3.703	< .001

and on how cohesive features can be used in conjunction with lexical features to explain human ratings of speech quality. The research has demonstrated that speech that contains more local cohesion devices (e.g., word repetition and causal cohesion) is generally rated to be of higher quality (Crossley & McNamara, 2013). Other research has demonstrated that speakers who integrate more words from the source text are judged to produce higher-quality speech samples (Crossley et al., 2014; Crossley & Kim, in press).

However, studies have not directly examined the links between source texts and speaking quality in terms of semantic similarity or keywords between the source text and the speaker's response. This feature is unique to TAACO and afforded us the opportunity to examine whether speaking samples that showed greater semantic overlap with the source texts were judged to be of higher quality. Such a finding would provide some evidence for intertextual cohesion, in which increased overlap between a source document and a response led to higher expert ratings, potentially because the increased overlap led to greater coherence on the part of the listener.

We found a significant but small effect for word2vec overlap between source and response. A full model including topic and TOEFL scores revealed a significant effect of key quadgrams, as well. These findings indicated that students who had greater similarity between their response and the source text, and who included in the response more key quadgrams taken from the source text, were judged to be more proficient speakers. Although the effect was small, it seems aligned with predictions based on integrated scoring rubrics. These rubrics mainly focus on speaking fluency, language use, and topic development. An element of topic development is conveying relevant information, which is likely what the source overlap indices were capturing. However, this is a relatively minor element of the grading rubric, which likely reflects the small amount of variance explained by the TAACO 2.0 indices. Indeed, much of the variance in scores was nonlinguistic, with conversations as source topics leading to higher scores, as well as with higher listening and structure scores on the TOEFL leading to higher scores.

Table 7 LME results for predicting TOEFL speaking score in Study 2

Fixed Effect (baseline)	Estimate	Std Error	*t*	*p*
(Intercept)	– 2.404	0.296	– 8.113	< .001
Topic: Notes (fungus)	0.453	0.060	7.537	< .010
Topic: Reciprocity (fungus)	0.138	0.068	2.021	< .001
Topic: Swimming (fungus)	0.219	0.070	3.137	< .050
Topic: Reciprocity (notes)	– 0.315	0.068	– 4.618	< .001
Topic: Swimming (notes)	– 0.234	0.067	– 3.497	< .001
Topic: Swimming (reciprocity)	0.082	0.047	1.725	> .050
TOEFL listening score	0.058	0.007	8.511	< .001
TOEFL structure score	0.014	0.006	2.250	< .050
Percentage key quad-grams that include a verb and/or a noun	1.200	0.500	2.401	< .050
Source similarity word2vec	0.901	0.170	5.292	< .001

The results from this analysis hold promise for better understanding not only integrated speaking tasks, but integrated written tasks as well. Integrated tasks are thought to be critical elements of academic success, because academic settings require students to read academic texts and listen to academic lectures and to cohesively integrate this information into class discussions as well as into written and oral reports (Douglas, 1997). However, researchers are just starting to understand the requirements of integrated tasks, even though the tasks have become common in a number of standardized testing situations (Cumming, Grant, Mulcahy-Ernt, & Powers, 2005a; Cumming et al., 2005b; Cumming, Kantor, Powers, Santos, & Taylor, 2000). The results reported here provide evidence linking source integration to measurements of success, likely because greater source similarity allows listeners to develop more coherent models of the speaking response. However, additional testing of this notion will be necessary, especially to parse out the relations between intertextual cohesion and rater coherence. Simple follow-up analyses could examine intertextual cohesion in integrated writing tasks. More optimal analyses would include human ratings of response cohesion and their links, to document overlap; think-aloud protocols, to examine whether intertextual cohesion influences expert raters; and behavioral studies that manipulate local, global, and intertextual cohesion, to examine their effects on text cohesion and rater coherence.

Conclusion

This article has introduced a new version of the Tool for the Automatic Analysis of Cohesion (TAACO 2.0). TAACO 2.0 reports on a number of indices related to local and global cohesion at the semantic level, as reported by latent semantic analysis, latent Dirichlet allocation, and word2vec. The tool also includes a source overlap feature that allows users to calculate lexical and semantic overlap between a source text and a response text (i.e., intertextual cohesion). We demonstrated the efficacy of the new tool in two studies, in order to test both the local and global indices based on semantic similarity indices, as well as the intertextual cohesion indices. Our findings indicate that the TAACO 2.0 indices are significant predictors of both text coherence and speaking quality. As with the original TAACO, we foresee TAACO 2.0 being used in a number of

prediction tasks that go beyond text coherence and speaking judgments. For instance, TAACO 1.0 has been used in a number of published studies already, in domains such as creativity (Skalicky, Crossley, McNamara, & Muldner, 2017), transcription disfluency (Medimorec, Young, & Risko, 2017), literary studies (Jacobs, Schuster, Xue, & Lüdtke, 2017), formative writing assessment (Wilson, Roscoe, & Ahmed, 2017), predicting math performance (Crossley, Liu, & McNamara, 2017a), self-regulated learning (Piotrkowicz et al., 2017), and medical discourse (Schillinger et al., 2017). We presume that researchers will continue to find innovative and discipline-specific applications of TAACO 2.0 in future research, especially considering the addition of the new semantic similarity metrics and source overlap scores.

We also plan to continuously update the TAACO tool as new approaches to measuring cohesion become available. In addition, we plan additional validation tests to make up for potential limitations in the present studies. For instance, we currently have only measured cohesion in relatively specialized corpora. We would like to expand the studies reported here to additional genres and registers. Additionally, we would like to examine cohesion in spoken data more directly, perhaps using human ratings of speech coherence as a gold standard.

Author note We thank Mary Jane White for early work on the data used in Study 1. Without her, this article would never have been possible. This project was supported in part by the National Science Foundation (DRL-1418378). We also thank YouJin Kim for her help collecting the data reported in Study 2. The ideas expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The research reported in Study 2 was funded by Educational Testing Service (ETS) under a Committee of Examiners and a Test of English as a Foreign Language research grant. ETS does not discount or endorse the methodology, results, implications, or opinions presented by the researchers. TOEFL test material is reprinted by permission of Educational Testing Service, the copyright owner. This research was also partially supported by Grant 644187 EC H2020 for the Realising an Applied Gaming Eco-System (RAGE) project, as well as by the European Funds of Regional Development through the Operation Productivity Program 2014–2020 Priority Axis 1, Action 1.2.1 D-2015, "Innovative Technology Hub Based on Semantic Models and High Performance Computing," Contract no. 6/1 09/2016.

Appendix: Holistic rating scale

After reading each essay and completing the analytical rating form, assign a holistic score based on the rubric below. For the following evaluations you will need to use a grading scale between 1 (minimum) and 6 (maximum). As with the analytical rating form, the distance between each grade (e.g., 1–2, 3–4, 4–5) should be considered equal.

SCORE OF 6: An essay in this category demonstrates clear and consistent mastery, although it may have a few minor errors. A typical essay effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position is well organized and clearly focused, demonstrating clear coherence and smooth progression of ideas exhibits skillful use of language, using a varied, accurate, and apt vocabulary demonstrates meaningful variety in sentence structure is free of most errors in grammar, usage, and mechanics.

SCORE OF 5: An essay in this category demonstrates reasonably consistent mastery, although it will have occasional errors or lapses in quality. A typical essay effectively develops a point of view on the issue and demonstrates strong critical thinking, generally using appropriate examples, reasons, and other evidence to support its position is well organized and focused, demonstrating coherence and progression of ideas exhibits facility in the use of language, using appropriate vocabulary demonstrates variety in sentence structure is generally free of most errors in grammar, usage, and mechanics.

SCORE OF 4: An essay in this category demonstrates adequate mastery, although it will have lapses in quality. A typical essay develops a point of view on the issue and demonstrates competent critical thinking, using adequate examples, reasons, and other evidence to support its position is generally organized and focused, demonstrating some coherence and progression of ideas exhibits adequate but inconsistent facility in the use of language, using generally appropriate vocabulary demonstrates some variety in sentence structure has some errors in grammar, usage, and mechanics.

SCORE OF 3: An essay in this category demonstrates developing mastery, and is marked by ONE OR MORE of the following weaknesses: develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence to support its position is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas displays developing facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice lacks variety or demonstrates problems in sentence structure contains an accumulation of errors in grammar, usage, and mechanics.

SCORE OF 2: An essay in this category demonstrates little mastery, and is flawed by ONE OR MORE of the following weaknesses: develops a point of view on the issue that is vague or seriously limited, and demonstrates weak critical thinking, providing inappropriate or insufficient examples, reasons, or other evidence to support its position is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas displays very little

facility in the use of language, using very limited vocabulary or incorrect word choice demonstrates frequent problems in sentence structure contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured.

SCORE OF 1: An essay in this category demonstrates very little or no mastery, and is severely flawed by ONE OR MORE of the following weaknesses: develops no viable point of view on the issue, or provides little or no evidence to support its position is disorganized or unfocused, resulting in a disjointed or incoherent essay displays fundamental errors in vocabulary demonstrates severe flaws in sentence structure contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning.

Holistic score based on attached rubric (1–6): ___

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*, 998–1006.

Bird, S., Klein, K., & Loper, E. (2009). Natural language processing with Python. Beijing, China: O'Reilly.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, *42*, 563–570.

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 740–750). Stroudsburg, PA: Association for Computational Linguistics.

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System*, *31*, 471–484. https://doi.org/10.1016/j.system.2003.02.002

Cree, G. S., & Armstrong, B. C. (2012). Computational models of semantic memory. In M. Spivey, K. McRae, & M. Joanisse (Eds.), The Cambridge handbook of psycholinguistics (pp. 259–282). New York, NY: Cambridge University Press.

Crossley, S. A., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*, 250–270.

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

Crossley, S. A., Liu, R., & McNamara, D. S. (2017a). Predicting math performance using natural language processing tools. In *LAK '17: Proceedings of the 7th International Learning Analytics and Knowledge Conference: Understanding, informing and improving learning with data* (pp. 339–347). New York, NY: ACM Press. https://doi.org/10.1145/3027385.3027399

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), Cognition in flux: Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (pp. 984–989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), Expanding the space of cognitive science: Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 1236–1241). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*, 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*, 171–192.

Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Grantee Submission*, *7*, 351–370.

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 *FLAIRS 2013—Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference* (pp. 208–213). Association for the Advancement of Artificial Intelligence.

Crossley, S. A., Russell, D., Kyle, K., & Römer, U. (2017b). Applying natural language processing tools to a student academic writing corpus: How large are disciplinary differences across science and engineering fields? *Journal of Writing Analytics*, *1*, 48–81.

Crossley, S. A., Weston, J. L., Sullivan, S. T. M., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*, 282–311. https://doi.org/10.1177/0741088311410188

Crowhurst, M. (1987). Cohesion in argument and narration at three grade levels. *Research in the Teaching of English*, *21*, 185–201.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005a). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL test (TOEFL Monograph No. MS-26). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & Jamse, M. (2005b). Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL® [ETS Research Report Series, 2005(1)]. Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). TOEFL 2000 writing framework. In *TOEFL-MS-18*. Princeton, NJ: Educational Testing Service.

Dascalu, M., McNamara, D. S., Trausan-Matu, S., & Allen, L. K. (2018). Cohesion network analysis of CSCL participation. *Behavior Research Methods*, *50*, 604–619. https://doi.org/10.3758/s13428-017-0888-4

Davies, M. (2008). The corpus of contemporary American English. Provo, UT: Brigham Young University.

Douglas, D. (1997). Testing speaking ability in academic contexts: Theoretical considerations. Princeton, NJ: Educational Testing Service.

Evola, J., Mamer, E., & Lentz, B. (1980). Discrete point versus global scoring of cohesive devices. In J. W. Oller & K. Perkins (Eds.), Research in language testing (pp. 177–181). Rowley, MA: Newbury House.

Faigley, L., & Witte, S. (1981). Analyzing Revision. *College Composition and Communication*, *32*, 400–414. https://doi.org/10.2307/356602

Fitzgerald, J., & Spiegel, D. L. (1983). Enhancing children's reading comprehension through instruction in narrative structure. *Journal of Reading Behavior*, *15*, 1–17. https://doi.org/10.1080/10862968309547480

Gernsbacher, M. A. (1990). Language comprehension as structure building. Hillsdale, NJ: Erlbaum.

Givón, T. (1995). Coherence in the text and coherence in the mind. In M. A. Gernsbacher & T. Givón, Coherence in spontaneous text (pp. 59–115). Amsterdam, The Netherlands: Benjamins.

Golub, G. H., & Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, *14*, 403–420.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet & C. E. Snow (Eds.), Rethinking reading comprehension (pp. 82–98). New York, NY: Guilford Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*, 193–202. https://doi.org/10.3758/BF03195564

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

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

Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication*, *17*, 307–352. https://doi.org/10.1177/0741088300017003001

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*, 346–363.

Jacobs, A. M., Schuster, S., Xue, S., & Lüdtke, J. (2017). What's in the brain that ink may character . . . : A quantitative narrative analysis of Shakespeare's 154 sonnets for use in neurocognitive poetics. *Scientific Study of Literature*, *7*, 4–51. https://doi.org/10.13140/RG.2.2.27126.40004

Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System*, *19*, 459–465. https://doi.org/10.1016/0346-251X(91)90026-L

King, M. L., & Rentel, V. (1979). Toward a theory of early writing development. *Research in the Teaching of English*, *13*, 243–253.

Kintsch, W. (1995). How readers construct situation models for stories: the role of syntactic cues and causal inferences. In M. A. Gernsbacher & T. Givón, Coherence in spontaneous text (pp. 139–160). Amsterdam, The Netherlands: Benjamins.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2004). Continuous multivariate distributions: Vol. 1. Models and applications. Hoboken, NJ: Wiley.

Kuznetsova, A., Brockhoff, B., & Christensen, H. B. (2016). lmerTest: Tests in linear mixed effects models (R package version 2.0-32). Retrieved from https://CRAN.R-project.org/package=lmerTest

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*, 319–340.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284. https://doi.org/10.1080/01638539809545028

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). Latent semantic analysis: A road to meaning. Mahwah, NJ: Erlbaum.

Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 2, pp. 302–308). Stroudsburg, PA: Association for Computational Linguistics.

Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, *33*, 623–636. https://doi.org/10.1016/j.system.2005.02.002

Loxterman, J. A., Beck, I. L., & McKeown, M. G. (1994). The effects of thinking aloud during reading on students' comprehension of more or less coherent text. *Reading Research Quarterly*, *29*, 353–367. https://doi.org/10.2307/747784

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning*, *60*, 501–533.

McCallum, A. K. (2002). *Mallet: A machine learning for language toolkit* (User's manual). Retrieved from http://mallet.cs.umass.edu/

McCulley, G. A. (1985). Writing quality, coherence, and cohesion. *Research in the Teaching of English*, *19*, 269–282.

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language*, *25*, 431–444. https://doi.org/10.1016/0749-596X(86)90036-7

McCutchen, D., & Perfetti, C. A. (1982). Coherence and connectedness in the development of discourse production. *Text—Interdisciplinary Journal for the Study of Discourse*, *2*, 113–140. https://doi.org/10.1515/text.1.1982.2.1-3.113

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

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499–515. https://doi.org/10.3758/s13428-012-0258-1

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–288. https://doi.org/10.1080/01638539609544975

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

Medimorec, S., Young, T. P., & Risko, E. F. (2017). Disfluency effects on lexical selection. *Cognition*, *158*, 28–32.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. ArXiv preprint. ArXiv:1310.4546

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*, 133–142.

Neuner, J. L. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English*, *21*, 92–105.

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "High-Stakes" measures of high school students' science achievement. *American Educational Research Journal*, *44*, 161–196.

Owen, S., Anil, R., Dunning, T., & Friedman, E. (2011). Mahout in action. Greenwich, CT, USA: Manning.

Piotrkowicz, A., Dimitrova, V., Treasure-Jones, T., Smithies, A., Harkin, P., Kirby, J., & Roberts, T. (2017). Quantified self analytics tools for self-regulated learning with myPAL. In *Proceedings of the 7th Workshop on Awareness and Reflection in Technology Enhanced Learning Co-located With the 12th European Conference on Technology Enhanced Learning (EC-TEL 2017)*. CEUR Workshop Proceedings. Retrieved from http://eprints.whiterose.ac.uk/121100/

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). https://doi.org/10.13140/2.1.2393.1847

Sanders, T., & Maat, H. P. (2006). Cohesion and coherence: Linguistic approaches. *Reading*, *99*, 440–466.

Schillinger, D., McNamara, D., Crossley, S., Lyles, C., Moffet, H. H., Sarkar, U., . . . Karter, A. J. (2017). The next frontier in communication and the ECLIPPSE study: Bridging the linguistic divide in secure messaging. *Journal of Diabetes Research*, *2017*, 1348242. https://doi.org/10.1155/2017/1348242

Skalicky, S., Crossley, S. A., McNamara, D. S., & Muldner, K. (2017). Identifying creativity during problem solving using linguistic features. *Creativity Research Journal*, *29*, 343–353.

Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing*, *34*, 16–36.

Yde, P., & Spoelders, M. (1985). Text cohesion: An exploratory study with beginning writers. *Applied Psycholinguistics*, *6*, 407–415. https://doi.org/10.1017/S0142716400006330