Reading and Writing (2023) 36:1053–1083
https://doi.org/10.1007/s11145-021-10221-x

Source inclusion in synthesis writing: an NLP approach to understanding argumentation, sourcing, and essay quality

Scott Crossley[1] · **Qian Wan**[1] · **Laura Allen**[1] · **Danielle McNamara**[1]

Accepted: 22 October 2021 / Published online: 11 November 2021

Abstract
Synthesis writing is widely taught across domains and serves as an important means of assessing writing ability, text comprehension, and content learning. Synthesis writing differs from other types of writing in terms of both cognitive and task demands because it requires writers to integrate information across source materials. However, little is known about how integration of source material may influence overall writing quality for synthesis tasks. This study examined approximately 900 source-based essays written in response to four different synthesis prompts which instructed writers to use information from the sources to illustrate and support their arguments and clearly indicate from which sources they were drawing (i.e., citation use). The essays were then scored by expert raters for holistic quality, argumentation, and source use/inferencing. Hand-crafted natural language processing (NLP) features and pre-existing NLP tools were used to examine semantic and keyword overlap between the essays and the source texts, plagiarism from the source texts, and instances of source citation and quoting. These variables along with text length and prompt were then used to predict essays scores. Results reported strong models for predicting human ratings that explained between 47 and 52% of the variance in scores. The results indicate that text length was the strongest predictor of score but also that more successful writers include stronger, semantically-related information from the source, provide more citations and do so later in the text, and copy less from the text. This work introduces the use of NLP techniques to assess source integration, provides details on the types of source integration used by writers, and highlights the effects of source integration on writing quality.

Keywords Synthesis writing · Natural language processing · Corpus linguistics

✉ Scott Crossley
scrossley@gsu.edu

[1] Department of Applied Linguistics/ESL, Georgia State University, Suite 1500, 25 Park Place, Atlanta, GA 30303, USA

Introduction

Synthesis writing is widely taught across academic disciplines and serves as an important means of assessing writing ability, text comprehension, and content learning (Vandermeulen et al., 2019). However, this form of writing can be extremely challenging, even for students in higher education (Van Ockenburg et al., 2018). Synthesis writing differs from other types of writing in terms of both cognitive and task demands because it requires writers to integrate information across source materials. Thus, it is a hybrid task that requires both reading and writing (Spivey & King, 1989); students must read and select appropriate information from sources, contrast and integrate that information while writing, and continuously revise texts as a result.

Synthesis writing is therefore unique because it focuses on selecting information from multiple texts, integrating, organizing, and connecting that information into a written text, and thematically structuring the information (Solé et al., 2013; Spivey, 1997; Spivey & King, 1989). A small number of studies have focused on the processes involved in information selection (Martínez et al., 2015; Mateos & Solé, 2009; Solé et al., 2013), whereas another line of research has sought to identify the product of integration (e.g., the amount, accuracy, and manner in which information is integrated into an essay from a source text, Gebril & Plakans, 2009, 2009; Uludag et al., 2019; Weigle & Parker, 2012). Importantly, product approaches have almost exclusively relied on hand annotations of source integration, which are costly in terms of the amount of time and money invested (Williamson et al., 2012).

The purpose of this study is to examine methods to automatically annotate argumentative synthesis essays for aspects of source integration. To do so, we annotate a large corpus of argumentative, synthesis essays composed by college students, military members, and the general adult population for aspects of source integration using natural language processing (NLP) techniques and examine links between these annotations and human ratings of writing quality. The NLP techniques we used include hand-crafted NLP features and pre-existing NLP tools that calculate features related to semantic and keyword overlap between the essays and the source texts, plagiarism from the source texts, and instances of source citation and quoting. The goal of the study is to examine the potential for these NLP features to predict human scores related to holistic essay quality and source use/inferencing beyond text length, which generally is a strong predictor of essay quality (Ferris, 1994; Frase, Faletti, Ginther, & Grant, 1999; Guo et al., 2013). Such an approach can provide information about the incidence of source integration features in source-based writing and the associations between these features and text quality as well as help assess the reliability of NLP features.

Argumentative synthesis writing

Argumentative synthesis writing prompts individuals to write essays using information from source texts that have been provided to them. Learning to write from and about sources is a key academic skill often developed during secondary and tertiary

education (Cumming, Lai, Cho, 2016; Vandermeulen et al, 2019). Synthesis tasks are often difficult, as they require the successful interpretation of the source material as well as the appropriate integration of information from the various sources (Martínez, Mateos, Martín, & Rijlaarsdam, 2015; Mateos et al., 2008; Solé et al., 2013). Despite the common use of source-based writing in academic contexts, many students struggle to master the writing task (Grabe & Zhang, 2013; Vandermeulen et al, 2019) even though the synthesis and integration of information into writing is considered essential for success, especially in tertiary education (Newell, Beach, & VanDerHeide, 2011; Haswell, 2000; Hirvela, 2011; Hood, 2008; Leki, 2017; Lillis & Curry, 2010; Melzer, 2009; Tardy, 2009). Indeed, success on synthesis tasks demonstrates that an individual is able to comprehend source texts, and in turn, develop written, argumentative responses that coherently incorporate source material (Solé et al., 2013; Spivey & King, 1989).

At its core, synthesis writing involves organizing ideas while reading and writing, selecting information from source texts, and integrating that information into writing based on inferences from the source texts (Spivey, 1997). More specifically, synthesis involves the accurate integration of source information, summarizing source information (Cumming, Lai, Cho, 2016), appropriate use of quotes, and avoiding plagiarism. Avoiding plagiarism during synthesis writing and developing ownership of information from a source is perhaps the most difficult task for writers beginning to practice synthesis writing. Thus, many studies have examined how students' text integration represents inappropriate or inadequate textual borrowing, assumedly resulting from lack of knowledge regarding cultural or discourse conventions (Bazerman, 2004; Belcher & Hirvela, 2001; Chandrasoma et al., 2004).

There are two general approaches to assessing successful synthesis in writing: process and product approaches. Process approaches often use audio-visual recordings (Martínez et al., 2015; Mateos & Solé, 2009; Solé et al., 2013) or keystroke logging (Leijten et al., 2019; Vandermeulen et al., 2019) to examine time spent reading sources and writing as well as transitions between the source texts and essay. Such studies can provide information about sub-processes involved in synthesis writing including comparing and contrasting information found in sources, linking sources, and integrating source information. A product approach, which is amenable to NLP techniques, focuses on identifying the amount of information integrated into an essay from a source text, how the information is integrated (e.g., quoting, paraphrasing, or summarizing), and the accuracy of the integration (Gebril & Plakans, 2009, 2009; Petrić, 2012; Plakans, 2009; Plakans & Gebril, 2013; Uludag et al., 2019; Weigle & Parker, 2012).

Source synthesis and writing proficiency

Research examining the product of successful synthesis writing generally follows two methods. The first method examines associations between features of source integration found in the essay and human scores of writing quality. The second method compares differences in source integration between first language (L1) and second language (L2) writers. The hypothesis underlying this approach is that L1

writers are more proficient than L2 writers and differences between the two populations can help identify features of successful writing.

Multiple studies have examined links between textual features related to source integration as found in argumentative essays and human ratings of quality. All of this work has relied on hand annotating features related to source integration and the majority of this research focused on L2 writing. For instance, studies have indicated that L2 writers with lower writing scores synthesized discourse to a lesser degree and used fewer source-based ideas in their essays (Plakans, 2009) that less accurately represented the source text (Uludag et al., 2019). Gebril and Plakans (2009) (2009 and Plakans and Gebril (2013) also reported that more advanced L2 writers used paraphrasing and summarization to integrate source material. Additionally, advanced L2 writers used direct source quoting to a greater extent than less advanced writers. Similarly, Petrić (2012) found that in L2 theses, higher scored writers use more direct quotations and that the quotations produced by higher proficiency writers were fragments while lower proficiency writers use clause-based quotations. In contrast, Weigle and Parker (2012) found that text integration in L2 writing was generally restricted to short excerpts from the source-text regardless of proficiency level and that there were few differences by proficiency level in how text was integrated into the essay, although they did report that less proficient students tended to quote more extensively. Leijten et al. (2019) also found that source integration was not a predictor of essay quality for L2 writers, but it was a predictor for L1 writers.

There are also reported differences in source integration between L1 and L2 writers during argumentative synthesis writing. For instance, studies have reported that L1 writers include more citations than L2 writers (Borg, 2000; van Weijen et al., 2019) and that citations that involve quotations are longer in L1 writing compared to L2 writing (Borg, 2000). Research also reports that when L2 writers do integrate text from the source, they seem to borrow more direct strings of words and are less likely to properly cite the source than L1 writers (Shi, 2004; van Weijen et al., 2019). More recently, van Weijen, Rijlaarsdam, and van den Bergh (2019) conducted a within-writer comparison of L1 and L2 source-based essays. The essays were annotated for six elements of source integration (e.g., unique source integration, source copying, and patchwriting). Van Weijen et al. found that source integration was more strongly related to person-level differences as compared to language differences (although L1 samples included more unique sources and less copied material than L2 samples). Most importantly, they also reported no clear effects of L2 proficiency on source integration.

Annotating source-based writing features

Annotating synthesis writing features (as in the product studies above) can provide important information about features important to source integration. However, as noted, most annotations of source integration have involved hand-coding. Common elements that have been hand-coded include indirect source integration (paraphrasing and summarizing), direct source integration with quoting, direct source integration without

quoting, total source integration (Gebril & Plakans, 2009, 2009; Plakans, 2009; Uludag et al., 2019), length of quoting, citation use (Leijten et al., 2019; Petrić, 2012; Weigle & Parker, 2012), switching between sources (Leijten et al., 2019), and source accuracy (Uludag et al., 2019). However, hand-coding data for language is a difficult, time-consuming, subjective, and resource intensive process (Dodigovic, 2005; Higgins et al., 2011) and is not a viable option for large corpora of texts. To analyze relevant linguistic features, patterns, and structures in large corpora, automated methods based on NLP are necessary (Granger et al., 2007; Meurers, 2015).

To date, most NLP approaches to source integration have focused on automatically detecting plagiarism. For instance, Clough and Stevenson (2011) examined algorithms to detect plagiarized texts (short passages) from Wikipedia articles. They developed new predictors of plagiarism including containment measures which calculated the normalized incidence of trigrams (i.e., three-word segments) that intersected between the source text and the short passages, and the longest sequence of words found in both the Wikipedia articles and the short passages (i.e., the longest common sequence). These features were able to accurately detect 80 percent of the plagiarized material. Other studies have examined how plagiarism detection can be improved by removing stop-words (e.g., pronouns, prepositions, and articles) and punctuations, lemmatizing words to bring them to their base form, predicting synonym use, and replacing numbers (Ceska & Fox, 2009; Chong et al., 2010).

A few studies have examined semantic similarity between source texts and written responses using word embedding algorithms such as Latent Semantic Analysis (LSA), which can compute semantic similarity between words, sentences, paragraphs, and texts based on distributional properties of words in large corpora. Studies have shown that semantic similarity between sources and responses is indicative of the quality of text summarizations and paraphrases (Crossley, Kyle, Davenport, & McNamara, 2016; Crossley et al., 2019).

Current study

The current study builds and extends previous research on synthesis writing by testing the predictive ability of NLP features to assess source integration. We extend previous NLP work on plagiarism and word embedding algorithms to assess the quality of source-based, argumentative essays in terms of holistic and source use/inferencing scores. We also introduce NLP features that represent aspects of source integration including features related to source citation and quoting. The research question that guides this study is: *To what degree can NLP features of source integration predict synthesis writing quality and source use/inferencing quality?*

Methods

Corpus

Our final corpus comprised 909 source-based essays collected from prior studies examining source-based writing.[1] The purpose of these experiments varied, from examining individual differences that contribute to source-based writing to interventions intended to improve the processing and integration of source information. Collectively, the essays were written by a diverse group of participants, including adults crowd-sourced from Mechanical Turk (n = 163), military recruits from the United States Navy (n = 177), and undergraduate college students from across the United States (n = 569). The Mechanical Turk participants were paid for their writing samples while the military members and the undergraduate students received course credit for participating. Demographic data was collected for all but 11 of the participants. The mean age of participants was 23.28 (SD = 8.73) with a minimum age of 17 and a maximum age of 74. The majority of the participants identified as White (n = 569) with the remaining participants identifying as Hispanic (n = 111), Asian (n = 86), Black (n = 83), Other (n = 46), Native-American (n = 2) or Middle Eastern (n = 1). In terms of gender, 401 of the participants were female and 497 were male. Two participants reported not graduating high-school and 245 reported having a high-school diploma or equivalent (but nothing else). The majority of the participants had some college experience (n = 550) with 61 reporting an Associate's degree. Forty-three participants reported obtaining a Bachelor's degree and one reported a Master's degree. Additionally, 25 participants reported being non-native speakers of English. In light of Van Weijin et al.'s (2019) findings that source integration patterns were linked more strongly to individual differences as opposed to language differences, we retained these participants, which in turn allowed us to keep the database intact.

The essays were written on one of four argumentative prompts related to *cell phone use and cancer*, *global warming*, *green living*, and the *locavore movement*. For each of these prompts, participants were given a set of 4–7 source documents that provided information about the topics. They were asked to read the sources and write a source-based essay; they were explicitly given instructions to elaborate on the information in the text instead of summarizing. They were also asked to use information from the texts to support their ideas, but to put ideas in their own words. Across the studies, participants were given around 40 min to read the sources and write an essay. Within the corpus, each participant wrote a single essay (i.e., there were 909 essays written by 909 individual participants). The average number of words written per essay was 355.92 (SD = 159.22) with a minimum number of words of 16 and maximum number of words of 1,313. Information for the essays in the corpus by prompt are provided in Table 1. Prompts and assignments are provided in "Appendix A".

[1] Our initial corpus was 919 texts. We removed ten texts because participants either did not write on topic or copy and pasted the entire essay from available sources.

Table 1 Descriptive statistics for corpus

Prompt	Number of essays	Number of sources
Cell phones	224	4
Global warming	215	4
Green living	314	6
Locavore movement	156	7

Table 2 Inter-rater reliability scores

Item	r	Kappa
Source use and inferencing	0.630	0.632
Holistic score	0.640	0.638

Scoring rubric

We developed a scoring rubric that included four analytic scores and one holistic score (see "Appendix B"). The analytic scores included argumentation, source use and inferencing, language sophistication, and organization. However, since this analysis was solely interested in the quality of source use, we did not analyze the argumentation, language sophistication or organization scores. Instead, we focus only on source use and inferencing scores and holistic scores.

For source use/inferencing, a good essay referred explicitly and accurately to the majority of sources, synthesized information across the sources, and went beyond simple paraphrasing of the sources. Thus, the source use and inferencing score refers to both source citation and source integration. The source use and inferencing scores were rated on a 1–4 scale (with 1 being low and 4 being high). Holistic essay score was based on a 1–6 scale with 1 labeled very poor and 6 labeled excellent.

Human raters

Human ratings for the analytic and holistic scores were provided by two expert raters. The raters were both white, female PhD students in an English composition program housed at a Predominantly Black Institution in the Southeastern United States. Both raters had 3 + years of teaching writing at the same university as well as 3 + years rating experience with standardized rubrics. The raters were paid $20 an hour for their work and were not included as co-authors on this paper. The raters were first trained on the rubric using source-based essays that were not part of the corpus used in this study. Raters scored each essay for analytic features first and then assigned a holistic score to the essay. When raters reached an acceptable level of reliability (Kappa > 0.70), they independently scored the source-based essays in the corpus. After initial scoring, Kappa scores ranged from 0.64 to 0.65 (see Table 2 for IRR scores). Raters then adjudicated all scores that showed a difference greater

Table 3 Human quality ratings overall and by prompt: Mean (SD)

Score	Overall	Cell phones	Global warming	Green living	Locavore
Source use and inferencing	2.83 (0.77)	2.87 (0.651)	2.93 (0.76)	2.92 (0.806)	2.45 (0.736)
Holistic	3.03 (1.02)	3.11 (0.878)	3.06 (1.060)	3.15 (1.110)	2.62 (0.844)

than one point. For these scores, raters discussed the essays and made adjustments as needed. The two scores for the raters were then averaged to create the final scores used in this study. After adjudication, all Kappa scores were greater than 0.7. Table 3 shows mean and standard deviation for source use/inferencing and holistic scores for all the essays and the essays by prompt. Source use/inferencing and holistic scores were correlated at $r = 0.734$.

Linguistic features

We used previously developed NLP measures of source integration along with bespoke NLP features to examine source integration and source citation in the source-based essays. We focused specifically on NLP feature sets that examine lexical and semantic overlap between the essay and the source texts, instances of plagiarism, source citation, and quoting. We discuss each of the NLP feature sets below. In total, we explored the use of 50 features. Descriptions of these features are provided in "Appendix C".

Essay and source overlap

We used the Tool for the Automatic Analysis of Cohesion (TAACO, Crossley et al., 2019) to examine keyword and semantic overlap between the essay and the source texts. In terms of keyword overlap, TAACO measures the degree to which key words and n-grams (i.e., single words and multi-word units) from the source texts are located in the target text. Key words and n-grams are identified by their relative frequency in the source texts compared to the frequency of the same items in the magazine and news sections of the Corpus of Contemporary American English (COCA, Davies, 2008). Once key words and n-grams are identified, TAACO calculates the proportion of these items in the essays written for each specific prompt. TAACO computes keyness indices for single words, bi-grams (two-word phrases), tri-grams (three-word phrases), and quad-grams (four-word phrases). TAACO also examines part of speech tags in n-grams for nouns, adjectives, and verbs wherein part of speech (POS) tags are allowed to substitute for words.

For semantic overlap, TAACO relies on LSA (Landauer et al., 1998), Latent Dirichlet Allocation (LDA, Blei et al., 2003), and Word2vec (Mikolov, Chen, Corrado, & Dean, 2013). Semantic spaces for each of these models were developed using the newspaper and magazine sections of COCA (Davies, 2008). Overlap between essay and source texts for each of these models was then computed to examine semantic similarity.

Plagiarism

We used the freely available Python package Plagiarism Detection (based on Chong, Specia, & Mitkov, 2010 and available at https://github.com/AashitaK/Plagiarism-Detection/blob/master/notebook.ipynb) to assess the longest common sequences of words shared between the source text and the essay. The Plagiarism Detection package also calculates a containment measurement score based on the intersection of tri-gram count in the source texts and tri-gram counts in the essay normalized by the number of tri-grams in the essay. It should be noted that the Plagiarism Detection package does not differentiate between quoted and unquoted text.

Source citation

We developed automated counts for citation use. Citations were defined as situations where the letters or numbers of the associate source texts, and/or the title, the author, the publisher of the texts, and the organization affiliated with the author, were mentioned in the writing. In this sample, the reference to the source text generally comprised the word "Source" and a single letter from "A–G" such as "Source A" or the parenthesized version of it, such as "(Source A)". Title, author, publisher, and organizational information were culled from the source text. We calculated simple citation metrics including citations counts, frequency of citation, percentage of most common cited source, source text coverage rate, and percent of paragraphs with citations. We also calculated locational information for citations including mean and standard deviation of location of citations in essays (both raw and normalized) by character location in essay, by sentence location, and by paragraph location. These variables allow us to calculate the depth of citations within an essay (e.g., are the citations found more near the beginning, middle, or end of an essay?).

Quoting

We developed automated counts for quotation use. Quotations were defined as any strings that were located between quotation marks in the essays. We calculated the number of quoted words in the text, the ratio of quoted words in the text, the number of quotations in the essay from the source text, and the percentage of quotations in the essay from the source texts.

Statistical analysis

The aim of this study was to investigate if NLP features related to source integration and citation were predictive of essays scores (source use/inferencing and holistic scores). We included text length as a baseline comparison because multiple studies have shown strong links between essay length and writing quality (Ferris, 1994; Frase, Faletti, Ginther, & Grant, 1999; Guo et al., 2013). We also included prompt

as a categorical factor to examine the potential for prompt-based effects (Crossley, Varner, & McNamara, 2013; Hinkel, 2002; Huot, 1990; Tedick, 1990). All statistical analyses were conducted in R 3.6.1.

Bivariate Pearson correlations were run using the cor.test() function (R core team, 2017) between our dependent variables (either human ratings of source use/inferences or holistic writing quality), text length, and the NLP features related to source integration and source citations. When NLP variables are highly collinear (i.e., potentially measuring the same feature) it can make interpreting variable importance difficult. Thus, prior to developing our models, we calculated correlations amongst the human ratings and the NLP variables (including text length). If two or more variables correlated at $r > 0.699$, the NLP variable(s) with the lowest correlation with the human scores was removed and the variable with the higher correlation was retained. We also only retained variables that demonstrated at least a small relationship with the response variable ($r > 0.099$).

We used the CARET package (Kuhn, 2016) in R (R Core Team, 2017) and the lm() function (R core team, 2017) to develop linear models. Model training and evaluation were performed using a stepwise tenfold cross-validation. For the stepwise process, we used the leapSeq function in Leaps (Kuhn et al., 2016). In the tenfold cross-validation procedure, the entire corpus was randomly divided into ten roughly equivalent sets and nine of these sets were used as a training set and one set was left out as test set. The model from the training set was then applied to the left-out test set. This happened ten times such that each set was used as the test set once. Estimates of accuracy are reported using average summary statistics across the ten test sets including root mean squared error (RMSE), mean absolute error (MAE) between the observed and modeled human scores, and the amount of variance explained by the developed model (R^2). It should be noted that since the source use/inferencing scores were scaled 1–4 and the holistic scores were scaled 1–6, direct comparisons between the models in terms of RMSE and MAE are not possible. F value and t values for each included variable were reported using lm() and the relative importance of the indices in each model was calculated using the calc.relimp() function in the relaimpo package (Grömping, 2009). Post-hoc analyses were conducted on the regression predictions for each model to see how they correlated with age differences in the populations and group differences (Mechanical Turk workers, Navy recruits, and undergraduate students).

Results

Source use and inferencing score

After controlling for multicollinearity amongst variables, 22 variables remained of which 14 showed at least a weak relationship with source use and inferencing scores. Of the 8 variables that did not report at least a small effect size, 7 of them were related to key words taken from the source text while the last variable was related to the number of times the most common source text was cited. The variables that demonstrated at least a small effect size were related to text length, source

Table 4 Descriptive statistics for selected variables

Variable	Mean	SD	Min	Max
Holistic score	3.03	1.02	0.25	6.00
Source use score	2.83	0.77	0.50	4.00
Number of words	355.92	159.22	14.00	1313.00
Standard deviation of citation location by word	65.56	55.57	0.00	294.45
Source coverage	0.49	0.30	0.00	1.00
Depth of citation location by word	152.58	110.15	0.00	595.80
Depth of citation location by character	783.10	660.72	0.00	3762.17
Citation count	3.67	2.78	0.00	18.00
Standard deviation citation location in paragraph (normed)	0.18	0.13	0.00	0.61
Standard deviation citation location in paragraph (raw)	1.55	1.62	0.00	9.45
Semantic overlap with source (word2vec)	0.88	0.07	0.00	0.98
Percentage of paragraphs with citations	0.59	0.35	0.00	1.00
Number of quotations from source	0.61	1.10	0.00	8.00
Depth of citation location by sentence (normed)	0.47	0.23	0.00	1.00
Number of quoted words	23.97	46.21	0.00	698.00
Semantic overlap with source (LDA)	0.99	0.02	0.83	1.00
Longest common subsequence	0.37	0.10	0.18	0.91

coverage, citation occurrence, semantic overlap between the essay and the sources, variance in citations, quotations use, and plagiarism (see Table 4 for descriptive statistics for each variable and Table 5 for correlation matrix for all variables and the source use/inferencing score).

When these 15 variables were entered into a tenfold cross-validated linear model, the number of variables that performed the best in explaining source use and inferencing score was 7, including number of words, the Locavore prompt, and variables related to source integration and source citations. The linear model reported RMSE = 0.558, MAE = 0.446, r = 0.687, $R^2 = 0.472$, F (9, 899) = 93.91, $p < 0.001$ (see model parameters summarized in Table 6). This model indicates that around 47% of the variance in the human scores for source use/inferencing can be explained by seven linguistic and experimental variables. The relative importance metrics indicate that the strongest predictor was number of words (explaining ~ 29% of the variance), followed by source coverage, depth of citations, semantic overlap between essay and source texts, variance in citation location, prompt, all of which were positive predictors. There was a single negative predictor, longest common subsequence (i.e., plagiarism), which reported the lowest importance. VIF values indicated no concerns with multicollinearity. Visual and statistical examinations of the residuals indicated they were normally distributed.

We conducted post-hoc analyses of the predicted source use/inferencing scores from the regression model to assess performance based on age and group (Mechanical Turk workers, Navy recruits, and undergraduate students). The correlation between age and predicted holistic score was nonsignificant ($r = 0.014$). A one-way ANOVA between groups reported statistical differences (F(2,

Table 5 Correlations between holistic score, source use/inferencing score, number of words, and source integration variables

Variable	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Holistic score	0.768	0.615	0.447	0.338	0.475	0.453	0.326	0.277	0.282	0.319	0.12	0.229	0.156	0.197	0.135	-0.321
2. Source use score	1	0.511	0.489	0.476	0.448	0.446	0.41	0.385	0.359	0.355	0.261	0.251	0.24	0.223	0.109	-0.187
3. Number of words		1	0.612	0.162	0.693	0.564	0.351	0.189	0.224	0.277	-0.06	0.274	0.065	0.327	0.169	-0.38
4. Standard deviation of citation location by word			1	0.471	0.672	0.572	0.523	0.478	0.447	0.251	0.296	0.229	0.293	0.34	0.118	-0.19
5. Source coverage				1	0.34	0.398	0.665	0.594	0.534	0.245	0.626	0.167	0.419	0.205	-0.049	0.009
6. Depth of citation location by word					1	0.665	0.51	0.323	0.283	0.272	0.159	0.227	0.539	0.299	0.188	-0.231
7. Depth of citation location by character						1	0.384	0.328	0.295	0.224	0.196	0.202	0.425	0.305	0.121	-0.162
8. Citation count							1	0.513	0.417	0.244	0.422	0.216	0.375	0.25	0.095	-0.061
9. Standard deviation citation location in paragraph (normed)								1	0.627	0.228	0.457	0.18	0.368	0.161	0.043	-0.068
10. Standard deviation citation location in paragraph (raw)									1	0.165	0.532	0.154	0.23	0.209	-0.041	-0.045
11. Semantic overlap with source (word2vec)										1	0.142	0.158	0.238	0.214	0.52	0.16
12. Percentage of paragraphs with citations											1	0.102	0.472	0.119	-0.092	0.016
13. Number of quotations from source												1	0.097	0.446	0.09	0.002
14. Depth of citation location by sentence (normed)													1	0.117	0.097	0.042
15. Number of quoted words														1	0.05	0.248
16. Semantic overlap with source (LDA)															1	0.055
17. Longest common subsequence																1

P values < .001 for *r* values > .109; p values < .050 for r values > .066

Table 6 Linear model to predict source use and inferencing score

Variable	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		2.880	0.039	74.866	< 0.001
Number of words	0.294	0.275	0.025	10.944	< 0.001
Source coverage	0.203	0.180	0.028	6.387	< 0.001
Prompt: Global warming	0.096[a]	0.049	0.055	0.892	> 0.050
Prompt: Green living	0.096[a]	0.005	0.053	0.103	> 0.050
Prompt: Locavore	0.096[a]	-0.385	0.063	-6.146	< 0.001
Semantic overlap with source (word2vec)	0.124	0.134	0.021	6.408	< 0.001
Standard deviation citation location in paragraph (normed)	0.102	0.065	0.024	2.742	< 0.010
Longest common subsequence	0.038	-0.058	0.021	-2.738	< 0.010
Depth of citation location by character	0.142	0.060	0.024	2.488	< 0.050

[a]Relative importance for prompt variable (not individual prompts)

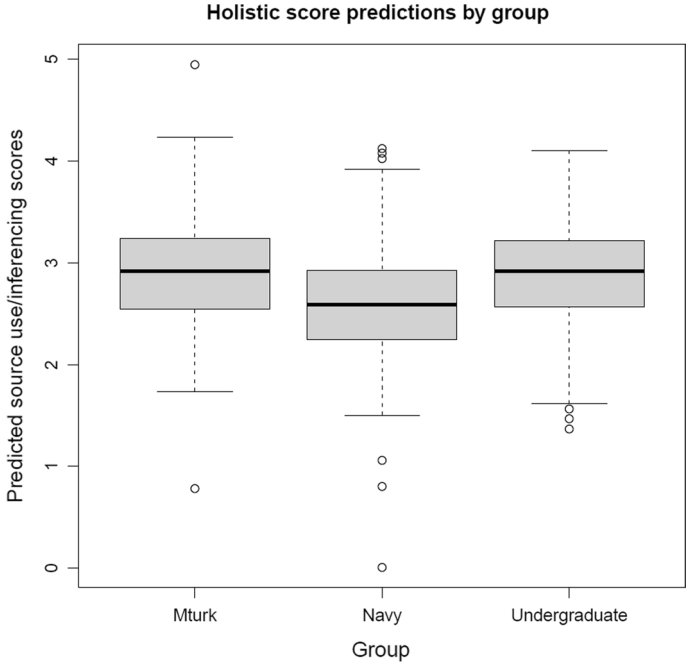

Fig. 1 Boxplots for source use/inferencing scores by group

906) = 23.39.82, $P < 0.001$). Tukey multiple comparisons indicated significant differences were reported between the predicted source use/inferencing scores for Navy recruits and Mechanical Turk workers and undergraduate students with

Table 7 Linear model to predict holistic score

Variable	Relative Importance	Estimate	SE	*t*	*p*
(Intercept)		3.243	0.049	65.853	< 0.001
Number of words	0.443	0.475	0.032	14.750	< 0.001
Prompt: Global warming	0.085[a]	-0.057	0.070	-0.811	< 0.001
Prompt: Green living	0.085[a]	-0.228	0.067	-3.415	> 0.050
Prompt: Locavore	0.085[a]	-0.705	0.080	-8.800	< 0.001
Source coverage	0.092	0.103	0.030	3.386	< 0.001
Longest common subsequence	0.122	-0.195	0.027	-7.195	< 0.001
Semantic overlap with source (word2vec)	0.103	0.170	0.027	6.340	< 0.010
Depth of citation location by character	0.155	0.098	0.031	3.145	< 0.010

[a] Relative importance for prompt variable (not individual prompts)

means indicating lower predicted source use/inferencing scores for Navy recruits (see Fig. 1 for a boxplot of data).

Holistic score

After controlling for multicollinearity amongst variables, 23 variables remained of which 15 showed at least a weak relationship with source use and inferencing scores. Of the 8 variables that did not report at least a small effect size, 7 of them were related to key words taken from the source text while the last variable was related to the number of times the most common source text was cited. The variables that demonstrated at least a small effect size were related to text length, depth of citation location, semantic overlap with sources, plagiarism, quoting, and source coverage (see Table 4 for descriptive statistics for each variable and Table 5 for a correlation matrix for all variables and holistic score). These were the same variables that were used in the source use and inferencing analysis.

When these 15 variables were entered into a tenfold cross-validated linear model, the number of variables that performed the best in explaining source use and inferencing score was 8, including number of words, two prompts, and variables related to source integration and source citations. The linear model reported RMSE = 0.715, MAE = 0.560, r = 0.720, $R^2 = 0.518$, F (8, 900) = 121.00, p < 0.001 (see model parameters summarized in Table 7). This model indicates that around 52% of the variance in the human scores for holistic essay quality can be explained by eight linguistic and experimental variables. The relative importance metrics indicate that the strongest, positive predictor was number of words (explaining ~ 44% of the variance) followed by depth of citations, semantic overlap between essay and source texts, and source coverage. There was one negative predictor, longest common subsequence (i.e., plagiarism), which explained about 12% of the variance. VIF values indicated no concerns with multicollinearity. Visual and statistical examinations of the residuals indicated they were normally distributed.

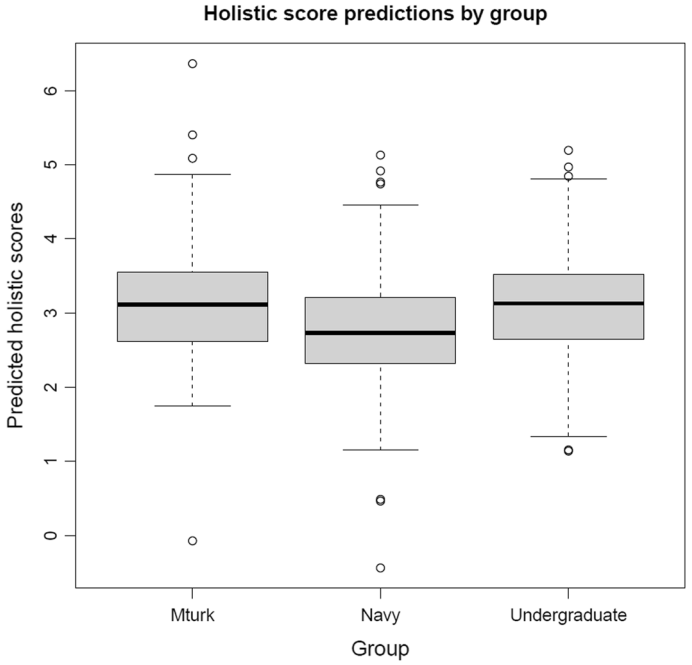

Fig. 2 Boxplots for holistic scores by group

We conducted post-hoc analyses of the predicted holistic scores from the regression model to assess performance based on non-textual features including age and group (Mechanical Turk workers, Navy recruits, and undergraduate students). The correlation between age and predicted holistic score was nonsignificant ($r = 0.006$). We used a one-way ANOVA between groups to examine differences in means. The ANOVA reported statistical differences ($F(2, 906) = 17.82$, $p < 0.001$). Tukey multiple comparisons indicated significant differences were reported between the predicted holistic scores for Navy recruits and Mechanical Turk workers and undergraduate students with means indicating lower predicted holistic scores for Navy recruits (see Fig. 2 for a boxplot of data).

Discussion

This study builds and extends previous research on assessing text integration in synthesis writing. Here, we introduce automated techniques to assess the products of synthesis writing including integrating source material (Cumming, Lai, Cho, 2016) and the use of citations, quotations, and plagiarism. The goal of this study was to examine how natural language processing (NLP) for annotating source integration and citations could be used to predict human ratings source use/inference and holistic scores when controlling for essay length and prompt. Our NLP techniques

focused on lexical and semantic similarity between sources and essays, plagiarism detectors, source citation metrics, and quoting.

The model for source use and inferencing explained ~ 47% of the variance in scores. As expected, the strongest predictor of the variance was number of words. However, five variables related to source integration and citation use explained variance beyond text length. The strongest predictor among these was source coverage, indicating that referring to a greater number of sources resulted in a higher score. Two other variables related to citation practice (i.e., standard deviation for citation location in paragraph and depth of citation location by character) were also significant predictors. These two variables indicated that stronger scores were related to producing citations later in the text (depth of citation) as well as spreading citations throughout the texts (standard deviation). Thus, it is not just incidence of citations that reflects good source use and inferencing; writers need to produce more sources later in the text, they need to cite a greater number of sources, and the citations need to be adequately spaced throughout the text.

A variable related to semantic overlap between the essay and the source text was also a strong predictor indicating that higher scored essays contained more semantic similarity with the source texts. In addition, more skilled writers are more likely to avoid copying long strings of words from the source text (i.e., plagiarism). Copying words from the source text is associated with lower source use and inferencing quality scores. The predictive features in this model align nicely with the expectations of the rubric, which indicated that greater source use and inferencing was the result of referring to a majority of sources, synthesizing information within and across sources, and providing interpretations of sources that were not simple iterations/paraphrases. As such, the results provide reliability metrics for these ratings, and in turn substantiate the claim that a number of NLP features can be used to provide proxies for certain aspects of source-based essays.

Our model of holistic score was also strong, predicting ~ 52% of the variance in the human ratings. As expected, the strongest predictor was text length, but four variables related to citation practices and source integration were also significant predictors beyond text length. These predictors were the same as found in the source use and inferencing scores, which indicate their importance in explaining writing quality in general (taking into consideration the relatively high correlation between holistic scores and source use/inferencing scores). The top predictor was related to depth of citation location (by character), indicating that greater holistic scores were related to citations being reported deeper in the essay. The next strongest predictor was a measure of plagiarism (i.e., longest common subsequence) followed by semantic overlap with sources and source coverage (i.e., the percentage of sources cited). In total, the model indicates that higher scored source-based essays result from the use of a greater number of citations that occur later in the text, fewer copied sequences from the source text, and greater lexical and topical overlap with the source, which were the same patterns reported in the source use and inference model.

We conducted post-hoc analyses for the regression predictions to better understand how the predictions differed amongst age levels and groups. The post-analyses for age for both the source use/inferencing scores and the holistic scores showed no associations. This result indicates that age seems to have no relationship with

the NLP operationalizations of source integration and citation practices and their links to human scores of writing quality. However, we did find differences in groups, such that Navy recruits' integration and citation practices and their effects on human scores of writing quality differed from college students and adult participants. Specifically, we found that Navy recruits used fewer integration and citation practices associated with successful writing when compared to other groups. These differences likely result from academic experiences and expectations between the Navy recruits and college freshmen and M-Turk populations. College freshmen are already enrolled in writing classes and have a strong motivation to successfully synthesize information into their writing as part of ongoing academic practice. M-Turk populations have also been shown to have higher education levels than the general population and are, thus, more likely to have experience with synthesis writing (Huff & Tingley, 2015).

We also noted differences between the source use/inferencing model and the holistic score model in terms of prediction accuracy. While both models included similar integration and citation features, the source use/inferencing model explained 47% of the variance while the holistic model explained 52% of the variance. At first glance, this may appear counterintuitive because the selected variables were generally associated with source integration, citation use, quotation practices, and plagiarism and one would expect these variables to predict source use and inferencing scores more strongly than holistic scores of writing. However, it is important to note that we included number of words as a variable in our models and that number of words was responsible for 44% of the holistic score model but only 29% of the source use and inferencing model. It is likely that the inclusion of number of words was the main difference in the total amount of variance explained between the two models.

Essay prompt was also a significant predictor in both the source use/inferencing and holistic models. In both models, lower scores were assigned to essays written on the Locavore prompt. This may the result of the Locavore prompt including seven different sources, the greatest number of sources in the data. For the holistic scores, the Global Warming prompt also showed significant differences, but it only contained four sources. Thus, it is unlikely that the number of sources per prompt influenced overall essay quality, but it may have influenced source use/inferencing scores which asked raters to consider the number of sources referenced.

The findings provide support for NLP approaches to not only explain successful source integration, but also for NLP approaches to provide automatic assessments of source integration that could be used to provide feedback to writers in online writing systems, allowing students opportunities at deliberate writing practice. From a researcher perspective, automatically annotating texts for evidence of source integration could also help spur additional research into source-integration studies. Previous studies that relied on hand coding of source integration features such as paraphrasing and summarizing, quoting, and citation use (Gebril & Plakans, 2009, 2009; Leijten et al., 2019; Petrić, 2012; Plakans, 2009; Uludag et al., 2019; Weigle & Parker, 2012) were resource intensive. The approaches used here can assess these features as well as additional synthesis features include plagiarism and citation depth, freeing researchers from hand-coding and allowing researchers to examine

larger corpora of essays from various student populations and academic disciplines. However, unlike hand annotations, our NLP approaches are unable to assess whether or not the text integration measured was successful. Our NLP approaches simply measure the existence of ideas and language from the sources (i.e., semantic overlap, topic overlap, quoting, plagiarism) and citation practices (depth of citation, range of citations, and variance in citations). The approaches cannot measure if the language or ideas integrated into an essay are accurate or if the citations used are meaningful.

The majority of the features investigated in this study were predictive of human scores of writing quality. For instance, after controlling for multicollinearity, 14 of the 22 remaining variables showed meaningful correlations with source use and inferencing scores (i.e., 64%). Of the 8 variables that did not report meaningful correlations, all but one calculated the inclusion of key word and POS n-grams from the source text. Similar results were reported for holistic scores in which 15 of the 23 variables (i.e., 65%) remaining after controlling for multicollinearity were meaningfully correlated with human scores. Like the source use and inferencing score, 7 of the 8 variables that were not correlated measured the inclusion of key word and POS n-grams from the source text. These finding indicate that the repetition of words and phrases from the source are not strongly associated with essay quality (although summarizing ideas from the source as reported by the semantic overlap scores was predictive). In terms of variables that were predictive of quality scores, it should be noted that quotations in general and specific quotations from the source text were not found to be predictive in the developed models, although they did report positive, but weak correlations with essay scores. Thus, while quoting text is associated with essay quality, it is likely not as important to modeling writing quality as other elements of source integration like location of citations, number of citations, source similarity, and plagiarism.

The models derived from these studies support previous research indicating that higher quality synthesis writing involves the accurate integration of source information (Cumming, Lai, Cho, 2016). For instance, like the L2 writers found in Plakans' studies (e.g., Gebril & Plakans, 2009, 2009; Plakans, 2009; Plakans & Gebril, 2013), we find that L1 writing that is of higher quality has greater overlap with source texts, indicating more semantic accuracy. However, unlike a few previous studies, we found no links between the use of quotes and the length of quotes and writing quality (Petrić, 2012), although we did find that borrowing longer strings of words from the source (i.e., plagiarism) did equate to lower writing quality (in a similar fashion to Shi, 2004; van Weijen et al., 2019). In terms of citation use, we find, like previous studies (i.e., Borg, 2000; van Weijen et al., 2019) that higher proficiency writers include more citations than lower proficiency writers.

Conclusion

Our initial research question for this study was: *To what degree can NLP features of source integration predict synthesis writing quality and source use/inferencing quality?* We found the variables related to source overlap, citation practices, and plagiarism explain a significant amount of variance in writing quality

beyond text length and prompt. In doing so, we introduced automatic approaches using NLP techniques to derive features from essays related to source integration. These include semantic and lexical overlap features between an essay and a source text, citation practices including number of citations, variance in citations, and position of citations in an essay, plagiarism features, and quoting features. A variety of these features were predictive of not only holistic essay scores but also source use and inferencing scores. The findings have important implications for source-based writing analyses, annotating source integration in texts, and source-based writing pedagogy.

We see this study as an initial investigation into the potential for automated NLP features of source integration to better explain source integration practices and link these practices with essay quality scores. While there are limitations to this study including sample size, participants, topics, and length of essays (some essays were shorter than 50 words), the study does provide strong evidence for the effectiveness of NLP features to examine source integration practices. However, there are many areas for development, especially considering that our models only explained around ~ 47% to ~ 52% of the variance in human ratings. We envision a number of different NLP metrics to increase our coverage of source integration in essay writings. For example, metrics for variation in sources use should be developed (i.e., what percentage of the overlap between the source and text comes from each source text and how much variation is there in that overlap). Additionally, future studies need to assess how well the NLP techniques used here are influenced by text length (i.e., are the results reliable with shorter texts). Beyond text length, it is not clear that the results would generalize to different populations (for instance middle or high school students), nor is it clear whether the results would hold across different writing prompts and sources. Thus, future studies that build on this work need to not only extend the NLP techniques used here but also the variety of texts, prompts, sources, and participants to ensure the results are generalizable.

Beyond NLP, there are a number of other techniques that could be used to identify other aspects of text integration. Eye-tracking could be used to examine time spent on reading sources, the number of sources read, and how often writers switch between sources. If combined with NLP approaches, eye-tracking techniques could tell us much about how reading patterns interact with actual source integration. Lastly, keystroke logging could complement NLP analyses and eye-tracking studies by examining writing bursts, which may be related to source integration, pauses, which may be related to reading of source texts, and determining copy-paste behaviors on the part of writers. Triangulating NLP, eye-tracking, and keystroke logging techniques may explain additional variance in human scores of source use and inferencing, and overall quality. A multidimensional account of source-based writing is needed to fully understand the complex interactions between comprehension and writing processes and how to scaffold students during the various stages of understanding multiple sources to help them convey their understanding of source texts within their essays.

Appendix A

See Table 8.

Table 8 Prompt and assignment information

Topic	Prompt
Cell phones	Write an essay that explains the effects of cell phones on humans and the extent to which cell phone use poses health risks Think carefully about the prompt. In your essay, elaborate on the information in the texts rather than merely summarizing. Please be as detailed as you can in your explanation Make sure to cite 3 or more sources in your essay. When you use information from the texts to support your essay, be sure to put ideas in your own words (e.g. paraphrasing, summarizing). Indicate clearly which sources you draw from in your essay You may cite sources by using the author's last name and year in parentheses (for example: Johnson, 2005) or as Source A, Source B, and so on The essay gives you an opportunity to show how effectively you can develop and express ideas. You should, therefore, take care to develop your point of view, present your ideas logically and clearly, and use language precisely
Global warming	Write an essay that explains the effects of climate change for life on earth and the extent to which humans are responsible Think carefully about the prompt. In your essay, elaborate on the information in the texts rather than merely summarizing. Please be as detailed as you can in your explanation Make sure to cite 3 or more sources in your essay. When you use information from the texts to support your essay, be sure to put ideas in your own words (e.g. paraphrasing, summarizing). Indicate clearly which sources you draw from in your essay You may cite sources by using the author's last name and year in parentheses (for example: Johnson, 2005) or as Source A, Source B, and so on The essay gives you an opportunity to show how effectively you can develop and express ideas. You should, therefore, take care to develop your point of view, present your ideas logically and clearly, and use language precisely
Green living	Green living (practices that promote the conservation and wise use of natural resources) has become a topic of discussion in many parts of the world today. With changes in the availability and cost of natural resources, many people are discussing whether conservation should be required of all citizens Carefully read the following six sources, including the introductory information for each source. Then synthesize information from at least three of the sources and incorporate it into a coherent, well-written essay that develops a position on the extent to which government should be responsible for fostering green practices Make sure that your argument is central, use the sources provided in the file links below to illustrate and support your reasoning. Avoid merely summarizing the sources. Indicate clearly which sources you are drawing from, whether through direct quotation, paraphrase or summary. You may cite sources as Source A, Source B, etc. or by using the descriptions in parentheses

Table 8 (continued)

Topic	Prompt
Locavore movement	Locavore are people who have decided to eat locally grown or produced products as much as possible. With an eye to nutrition as well as sustainability (resource use that preserves the environment), the locavore movement has become widespread over the past decade Imagine that a community is considering organizing a locavore movement. Carefully read the following seven sources, including the introductory information for each source. Then synthesize information from at least three of the sources and incorporate it into a coherent, well-developed essay that identifies the key issues associated with the locavore movement ad examines their implications for the community Make sure that your argument is central, use the sources provided in the file links below to illustrate and support your reasoning. Avoid merely summarizing the sources. Indicate clearly which sources you are drawing from, whether through direct quotation, paraphrase or summary. You may cite sources as Source A, Source B, etc. or by using the descriptions in parentheses

Appendix B

See Table 9.

Table 9 Integrated essay scoring guidelines

	Argumentation	Source use and inferencing	Language sophistication	Organization
4	The essay, in general: Discusses the side(s) of the argument and explicitly states a position Supports the position by providing 3 + relevant and accurate claims Supports the position and claims by providing 3 + relevant and accurate pieces of evidence	The essay, in general: Explicitly and accurately refers to the majority of outside sources Synthesizes information from both within and across the referenced sources Provides deep interpretations of sources that go far beyond simple reiteration/paraphrase of the material	The essay, in general: Demonstrates lexical and syntactic sophistication and variety in word choice and syntax Uses consistently appropriate word choices Demonstrates a command of English spelling, grammar, and mechanics, containing few to no errors	The essay, in general: Follows a logical structure, beginning with an introduction and ending with concluding statements Is well-organized and maintains sense of flow throughout the paragraphs Is coherent and makes appropriate use of cohesive devices to signal connections between ideas
3	The essay, in general: Discusses the side(s) of the argument and implicitly states a position Supports the position by providing 1–2 relevant and accurate claims Supports the position and claims by providing 1–2 relevant and accurate pieces of evidence	The essay, in general: Explicitly and accurately refers to one or more of the outside sources Generates explicit connections from information within sources but generates few explicit or implicit connections across the multiple sources Provides interpretations of sources that go somewhat beyond reiteration/paraphrase of the material	The essay, in general: Demonstrates lexical and syntactic sophistication but little variety in word choice and syntax Generally uses appropriate word choices Shows an understanding of English spelling, grammar, and mechanics, but may contain some errors	The essay, in general: Follows a logical structure but lacks explicit introduction or concluding statements Contains evidence of organization but may lack some appropriate transitions between paragraphs Is coherent and generally makes appropriate use of cohesive devices to signal connections between ideas
2	The essay, in general: Discusses the side(s) of the argument but does not provide a position Discusses the side(s) by providing 1 + relevant or accurate claims Supports the side(s) by providing 1 + relevant or accurate pieces of evidence	The essay, in general: Implicitly refers to one or more of the outside sources Generates explicit or implicit connections from information within sources but fails to generate any connections across the multiple sources Relies heavily on direct quotes or paraphrases of the source material	The essay, in general: Demonstrates little lexical and syntactic sophistication and little variety Contains some inappropriate word choices Shows some understanding of English spelling, grammar, and mechanics, but contains numerous errors	The essay, in general: Sometimes deviates from logical structure and lacks introduction or concluding statements Contains some evidence of organization but lacks important transitions between central ideas and paragraphs Is somewhat coherent but lacks important cohesive elements

Table 9 (continued)

	Argumentation	Source use and inferencing	Language sophistication	Organization
1	The essay, in general: Does not discuss the side(s) of the argument nor provide a position Provides no claims to support a position Provides no evidence to support the side(s) or position	The essay, in general: Does not refer to any of the provided sources Does not synthesize information from within or across the sources provided Fails to reference concepts described in the source material	The essay, in general: Demonstrates low lexical and syntactic sophistication and little to no variety Generally does not use appropriate word choices Contains a number of spelling, grammar, and mechanics errors that render portions of the text difficult to understand	The essay, in general: Lacks a logical sequence of thought Lacks an appropriate organizational structure Is not coherent and lacks important cohesive elements

Holistic Essay Quality: _____ / 6

1-Very poor

2-Poor

3-Fair

4-Good

5-Very good

6-Excellent

Appendix C

See Table 10.

Table 10 NLP features used in modeling

Name	Description	Construct	Calculation
Percentage key unigrams	Percentage of unigrams in text that are keywords	Essay and Source Overlap	TAACO
Percentage key unigrams (nouns)	Percentage of unigrams in text that are keywords (nouns)	Essay and Source Overlap	TAACO
Percentage key unigrams (verbs)	Percentage of unigrams in text that are keywords (verbs)	Essay and Source Overlap	TAACO
Percentage key unigrams (verbs and nouns)	Percentage of unigrams in text that are keywords (verbs and nouns)	Essay and Source Overlap	TAACO
Percentage key unigrams (adjectives)	Percentage of unigrams in text that are keywords (adjectives)	Essay and Source Overlap	TAACO
Percentage key bigrams	Percentage of bigrams in text that are keywords (all bigrams)	Essay and Source Overlap	TAACO
Percentage key trigrams	Percentage of trigrams in text that are keywords (all trigrams)	Essay and Source Overlap	TAACO
Percentage key quadgrams	Percentage of quadgrams in text that are keywords (all quadgrams)	Essay and Source Overlap	TAACO
Percentage key bigrams (nouns)	Percentage of bigrams in text that are keywords (bigrams that include a noun)	Essay and Source Overlap	TAACO
Percentage key bigrams (adjectives)	Percentage of bigrams in text that are keywords (bigrams that include an adjective)	Essay and Source Overlap	TAACO
Percentage key bigrams (verbs)	Percentage of bigrams in text that are keywords (bigrams that include a verb)	Essay and Source Overlap	TAACO
Percentage key bigrams (verbs and nouns)	Percentage of bigrams in text that are keywords (bigrams that include a verb and/or a noun)	Essay and Source Overlap	TAACO
Percentage key bigrams (adjectives and nouns)	Percentage of bigrams in text that are keywords (bigrams that include an adjective and/or a noun)	Essay and Source Overlap	TAACO
Percentage key trigrams (nouns)	Percentage of trigrams in text that are keywords (trigrams that include a noun)	Essay and Source Overlap	TAACO
Percentage key trigrams (adjectives)	Percentage of trigrams in text that are keywords (trigrams that include an adjective)	Essay and Source Overlap	TAACO

Table 10 (continued)

Name	Description	Construct	Calculation
Percentage key trigrams (verbs)	Percentage of trigrams in text that are keywords (trigrams that include a verb)	Essay and Source Overlap	TAACO
Percentage key trigrams (verbs and nouns)	Percentage of trigrams in text that are keywords (trigrams that include a verb and/or a noun)	Essay and Source Overlap	TAACO
Percentage key trigrams (adjectives and nouns)	Percentage of trigrams in text that are keywords (trigrams that include an adjective and/or a noun)	Essay and Source Overlap	TAACO
Percentage key quadgrams (nouns)	Percentage of quadgrams in text that are keywords (quadgrams that include a noun)	Essay and Source Overlap	TAACO
Percentage key quadgrams (adjectives)	Percentage of quadgrams in text that are keywords (quadgrams that include an adjective)	Essay and Source Overlap	TAACO
Percentage key quadgrams (verbs)	Percentage of quadgrams in text that are keywords (quadgrams that include a verb)	Essay and Source Overlap	TAACO
Percentage key quadgrams (verbs and nouns)	Percentage of quadgrams in text that are keywords (quadgrams that include a verb and/or a noun)	Essay and Source Overlap	TAACO
Percentage key quadgrams (adjectives and nouns)	Percentage of quadgrams in text that are keywords (quadgrams that include an adjective and/or a noun)	Essay and Source Overlap	TAACO
Semantic overlap with source (LSA)	Latent semantic cosine similarity between target text and source text	Essay and Source Overlap	TAACO
Semantic overlap with source (LDA)	Latent dirichlet allocation divergence score between target text and source text	Essay and Source Overlap	TAACO
Semantic overlap with source (word2vec)	Word2vec similarity score between target text and source text	Essay and Source Overlap	TAACO
Longest common subsequence	The longest sequence of words included in both source texts and essay	Plagiarism	Plagiarism Detection tool
Raw Containment	Intersection of tri-gram count in the source texts and essay (raw count)	Plagiarism	Plagiarism Detection tool

Table 10 (continued)

Name	Description	Construct	Calculation
Average containment	Intersection of tri-gram count in the source texts and essay normalized by the number of tri-grams in the essay	Plagiarism	Plagiarism Detection tool
Number of quoted words	Raw count of quoted words	Quoting	Internal
Number of quotations from source	Number of quotes taken from source	Quoting	Internal
Percentage quotations from source	The number of quotes taken from source divided by total number of quotes in the essay	Quoting	Internal
Ratio of quoted words	Ratio of quoted words to number of words in texts	Quoting	Internal
Depth of citation location by character	Average character-based location of all citation instances. For each citation instance, the index of its first character was recorded as the character-based location	Source citation	Internal
Depth of citation location by sentence (normed)	Average normalized sentence-based location of all citation instances. For each citation instance, the normalized sentence-based location is the index of the sentence that contains the citation instance divided by the total number of sentences in the essay	Source citation	Internal
Depth of citation location in paragraph by sentence (normed)	Average normalized sentence-based location (in paragraph) of all citation instances. For each citation instance, the normalized sentence-based location in paragraph is the index of the sentence that contains the citation instance divided by the total number of sentences in the paragraph	Source citation	Internal
Depth of citation location by sentence	Average sentence-based location of all citation instances. For each citation instance, the index of the sentence that contains the citation instance was recorded as the sentence-based location	Source citation	Internal

Table 10 (continued)

Name	Description	Construct	Calculation
Depth of citation location in paragraph by sentence (raw)	Average sentence-based location (in paragraph) of all citation instances. For each citation instance, the index of the sentence that contains the citation instance within the paragraph was recorded as the sentence-based location in paragraph	Source citation	Internal
Depth of citation location by word	Average word-based location of all citation instances. For each citation instance, the index of its first word was recorded as the word-based location	Source citation	Internal
Citation count	Raw count of citations	Source citation	Internal
Frequency of citations	Normed count of citations (normed by word count)	Source citation	Internal
Percent most common cited source	The count of citations to the most commonly cited source (by the essay writer) divided by the total number of citations in the essay	Source citation	Internal
Percentage of paragraphs with citations	The number of paragraphs that contained at least one citation divided by the total number of paragraphs in an essay	Source citation	Internal
Standard deviation of citation location by character	Standard deviation of character-based locations of all citation instances for essays that contain more than one citation instance, otherwise 0 was used as a holder value	Source citation	Internal
Standard deviation norm sentence location in essay	Standard deviation of normalized sentence-based locations of all citation instances for essays that contain more than one citation instance, otherwise 0 was used as a holder value	Source citation	Internal
Standard deviation citation location in paragraph (normed)	Standard deviation of normalized sentence-based locations (in paragraph) of all citation instances for essays that contain more than one citation instance, otherwise 0 was used as a holder value	Source citation	Internal

Table 10 (continued)

Name	Description	Construct	Calculation
sd raw sentence location in essay	Standard deviation of sentence-based locations of all citation instances for essays that contain more than one citation instance, otherwise 0 was used as a holder value	Source citation	Internal
Standard deviation citation location in paragraph (raw)	Standard deviation of sentence-based locations (in paragraph) of all citation instances for essays that contain more than one citation instance, otherwise 0 was used as a holder value	Source citation	Internal
Standard deviation of citation location by word	Standard deviation of word-based locations of all citation instances for essays that contain more than one citation instance, otherwise 0 was used as a holder value	Source citation	Internal
Source coverage	The ratio of source texts cited to the number of all source texts available	Source citation	Internal

Acknowledgements This research was supported in part by the Institute for Education Sciences (IES R305A180261 and R305A180144) and the Office of Naval Research (N00014-20-1-2623). The deas expressed in this material are those of the authors and do not necessarily reflect the views of our funders.

References

AashitaK/Plagiarism-Detection. (n.d.). GitHub. Retrieved May 19, 2021, from https://github.com/AashitaK/Plagiarism-Detection/blob/master/notebook.ipynb.

Bazerman, C. (2004). Intertextuality: How texts rely on other texts. In Bazerman, C., & Prior, P. (Eds.), *What writing does and how it does it: An introduction to analyzing texts and textual practices* (1st ed., pp. 83–96). Routledge. https://doi.org/10.4324/9781410609526

Belcher, D., & Hirvela, A. (Eds.). (2001). Linking Literacies. University of Michigan Press. https://doi.org/10.3998/mpub.11496

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Borg, E. (2000). Citation practices in academic writing. In Thompson, P. (Ed.), *Patterns and perspectives: Insights into EAP writing practice* (pp. 26–42). Reading, UK: Centre for Applied Language Studies.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Ceska, Z., & Fox, C. (2009). The Influence of Text Pre-processing on Plagiarism Detection. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., & Nikolov, N. (Eds.), *Proceedings of the International Conference RANLP-2009* (pp. 55–59). Association for Computational Linguistics.

Chandrasoma, R., Thompson, C., & Pennycook, A. (2004). Beyond Plagiarism: Transgressive and nontransgressive intertextuality. *Journal of Language, Identity & Education, 3*(3), 171–193. https://doi.org/10.1207/s15327701jlie0303_1

Chong, M., Specia, L., & Mitkov, R. (2010). Using natural language processing for automatic detection of plagiarism. In *Proceedings of the 4th International Plagiarism Conference* (IPC-2010).

Clough, P., & Stevenson, M. (2011). Developing a corpus of plagiarised short answers. *Language Resources and Evaluation, 45*(1), 5–24. https://doi.org/10.1007/s10579-009-9112-1

Crossley, S. A, Kyle, K., Davenport, J., & McNamara, D. S. (2016). Automatic assessment of constructed response data in a chemistry tutor. In Barnes, T., Chi, M., & Feng, M. (eds.). *Proceedings of the 9th International Educational Data Mining* (pp. 336–340). EDM Society.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods, 51*, 14–27. https://doi.org/10.3758/s13428-018-1142-4

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

Cumming, A., Lai, C., & Cho, H. (2016). Students' writing from sources for academic purposes: A synthesis of recent research. *Journal of English for Academic Purposes, 23*, 47–58. https://doi.org/10.1016/j.jeap.2016.06.002

Davies, M. (2008). The Corpus of Contemporary American English. www.english-corpora.org/coca/

Dodigovic, M. (2005). *Artificial intelligence in second language learning: Raising error awareness*. Multilingual Matters.

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

Frase, L., Faletti, J., Ginther, A., & Grant, L. (1999). Computer analysis of the TOEFL test of written English (TOEFL Research Report No. 64). Princeton, NJ: ETS.

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7(1)*, 47–84.

Grabe, W., & Zhang, C. (2013). Reading and writing together: A critical component of English for academic purposes teaching and learning. *TESOL Journal, 4*(1), 9–24. https://doi.org/10.1002/tesj.65

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL, 19*(3), 252–268. https://doi.org/10.1017/s0958344007000237

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician, 63*(4), 308–319. https://doi.org/10.1198/tast.2009.08199

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*(3), 218–238.

Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication, 17*(3), 307–352. https://doi.org/10.1177/0741088300017003001

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language, 25*(2), 282–306. https://doi.org/10.1016/j.csl.2010.06.001

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

Hirvela, A. (2011). Writing to learn in content areas: Research insights. In ManchónR. M. (Ed.), *Learning-to-Write and Writing-to-Learn in an Additional Language* (pp. 37–59). Amsterdam/Philadelphia John Benjamins Publishing Company.

Hood, S. (2008). Summary writing in academic contexts: Implicating meaning in processes of change. *Linguistics and Education, 19*(4), 351–365. https://doi.org/10.1016/j.linged.2008.06.003

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics, 2*(3), 2053168015604648.

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

Kuhn, K. (2016). Contributions from Wing J, Weston S, Williams A, Keefer C, Engelhardt A, et al. caret: Classification and Regression Training. R package version, 6-0.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Leki, I. (2017). *Undergraduates in a second language: Challenges and complexities of academic literacy development*. Routledge.

Leijten, M., Van Waes, L., Schrijver, I., Bernolet, S., & Vangehuchten, L. (2019). Mapping master's students'use of external sources in source-based writing in L1 and L2. *Studies in Second Language Acquisition, 41*(3), 555–582. https://doi.org/10.1017/s0272263119000251

Lillis, T. M., & Curry, M. J. (2010). *Academic writing in global context*. Routledge.

Martínez, I., Mateos, M., Martín, E., & Rijlaarsdam, G. (2015). Learning history by composing synthesis texts: Effects of an instructional programme on learning, reading and writing processes, and text quality. *Journal of Writing Research, 7*(2), 275–302. https://doi.org/10.17239/jowr-2015.07.02.03

Mateos, M., Martín, E., Villalón, R., & Luna, M. (2008). Reading and writing to learn in secondary education: Online processing activity and written products in summarizing and synthesizing tasks. *Reading and Writing, 21*(7), 675–697. https://doi.org/10.1007/s11145-007-9086-6

Mateos, M., & Solé, I. (2009). Synthesising information from various texts: A study of procedures and products at different educational levels. *European Journal of Psychology of Education, 24*(4), 435–451. https://doi.org/10.1007/bf03178760

Melzer, D. (2009). Writing assignments across the curriculum: A national study of college writing. *College Composition and Communication, 61*(2), W240–W261.

Meurers, D. (2015). Learner corpora and natural language processing. In Granger, S., Gaëtanelle Gilquin, & Meunier, F. (Eds.), *The Cambridge handbook of learner corpus research* (pp. 537–566). Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Newell, G. E., Beach, R., Smith, J., & VanDerHeide, J. (2011). Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly, 46*(3), 273–304. https://doi.org/10.1598/RRQ.46.3.4

Ockenburg, L. V., Weijen, D. V., & Rijlaarsdam, G. (2018). Syntheseteksten leren schrijven in het voortgezet onderwijs; Het verband tussen schrijfaanpak en voorkeur voor leeractiviteiten. *Levende Talen Tijdschrift, 19*(2), 3–14.

Petrić, B. (2012). Legitimate textual borrowing: Direct quotation in L2 student writing. *Journal of Second Language Writing, 21*(2), 102–117. https://doi.org/10.1016/j.jslw.2012.03.005

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26*(4), 561–587. https://doi.org/10.1177/0265532209340192

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing, 17*(1), 18–34. https://doi.org/10.1016/j.asw.2011.09.002

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(3), 217–230. https://doi.org/10.1016/j.jslw.2013.02.003

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

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*(2), 171–200. https://doi.org/10.1177/0741088303262846

Solé, I., Miras, M., Castells, N., Espino, S., & Minguela, M. (2013). Integrating information: An analysis of the processes involved and the products generated in a written synthesis task. *Written Communication, 30*(1), 63–90. https://doi.org/10.1177/0741088312466532

Spivey, N. (1997). *The constructivist metaphor: Reading, writing, and making of meaning*. Academic Press.

Spivey, N. N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Research Quarterly, 24*(1), 7–26. https://doi.org/10.1598/rrq.24.1.1

Tardy, C. M. (2009). *Building genre knowledge*. Parlor Press.

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

Uludag, P., Lindberg, R., McDonough, K., & Payant, C. (2019). Exploring L2 writers' source-text use in an integrated writing assessment. *Journal of Second Language Writing, 46*, 100670. https://doi.org/10.1016/j.jslw.2019.100670

Vandermeulen, N., van den Broek, B., Van Steendam, E., & Rijlaarsdam, G. (2019). In search of an effective source use pattern for writing argumentative and informative synthesis texts. *Reading and Writing, 33*(2), 239–266. https://doi.org/10.1007/s11145-019-09958-3

van Weijen, D., Rijlaarsdam, G., & van den Bergh, H. (2019). Source use and argumentation behavior in L1 and L2 writing: A within-writer comparison. *Reading and Writing, 32*(6), 1635–1655. https://doi.org/10.1007/s11145-018-9842-9

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing, 21*(2), 118–133. https://doi.org/10.1016/j.jslw.2012.03.004

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x