Argumentation Features and Essay Quality: Exploring Relationships and Incidence Counts

Scott A. Crossley, Yu Tian & Qian Wan

Department of Applied Linguistics/ESL, Georgia State University | United States

Abstract: This study examines links between human ratings of writing quality and the incidence of argumentative features (e.g., claims, data) in persuasive essays along with relationships among these features and their distance from one another within an essay. The goal is to better understand how argumentation elements in persuasive essays combine to model human ratings of essay quality. The study finds that, in most cases, it is not the presence of argumentation features that is predictive of writing quality but rather the relationships between superordinate and subordinate features, parallel features, and the distances between features. This finding has not only theoretical value but also practical value in terms of pedagogical approaches and automated writing feedback.

Keywords: argumentation, rhetorical structure theory, writing assessment

Crossley, S. A., Tian, Y., & Wan, Q. (2022). Argumentation features and essay quality: Exploring relationships and incidence counts. *Journal of Writing Research*, *14*(1), 1-34 - https://doi.org/10.17239/jowr-2022.14.01.01

Contact: Scott A. Crossley, Department of Applied Linguistics/ESL, Georgia State University, 25 Park Place, Suite 1500, Atlanta, GA 30303 | United States – scrossley@gsu.edu.

1. Introduction

Argumentative writing is a crucial skill in personal, professional, and academic contexts (Lee & Deakin, 2016; Pessoa, Mitchell, & Miller, 2017). In academic settings, the ability to construct persuasive arguments that express point of view using appropriate academic language is essential not only when writing generic, context-neutral academic essays for writing courses, but also when writing within different disciplines (Hirvela, 2017; Varghese & Abraham, 1998). Indeed, argumentative writing is particularly cognitively demanding because it entails a hierarchical, analytic structure that requires critical arguments to be systematically supported (Applebee, 1984). This feature makes argumentative writing one of the most difficult types of writing to produce (Berrill,1996; Crasnich & Lumbelli, 2005; Gárate & Melero, 2005) and tertiary students often reach university with different writing experiences and levels of exposure to argumentation types. While university students are expected to adequately engage in written argumentation (Currie, 1996), many of them lack sufficient understanding of what constitutes robust argumentation and how different argumentation elements are organized to present a reasoned argument in their writing (Hyland, 1997).

There exist a variety of approaches to analyze and annotate argumentative structures in essays to better understand their form and structure. These include approaches that classify segments of writing into specific argumentation features including claims and data (e.g., Toulmin, 1958) and approaches that examine relationships between argument features (Azar, 1999; Ferretti, Lewis, & Andrews-Weckerly, 2009). Annotating essays for argumentative features allows us to better understand argumentative strategies and how the use of argumentative features leads to successful writing. In the case of classifying argumentative features, previous research (Nussbaum & Kardash, 2005; Qin & Karabacak, 2010; Stapleton & Wu, 2015) has examined correlations between the incidences of these features and holistic scores of writing quality, generally finding that a greater number of argumentative features leads to stronger writing scores. However, no research to our knowledge has examined how relationships between argumentative features such as final claims, primary claims, data along with simple annotations of these argument features can model human ratings of essay quality. Such an approach might demonstrate the importance of argumentative features in persuasive writing and allow for investigations of how these elements interact in modeling human ratings of writing quality.

1.1 Argumentation Schema

In writing, the process of argumentation is generally formalized with identifiable elements that can be isolated and analyzed (Stapleton & Wu, 2015). Considerable efforts have been made to devise argumentation schemes to facilitate the

categorization and analysis of these elements (Walton, Reed, & Macagno, 2008), among which Toulmin's (1958, 2003) scheme has been widely recognized as a universal system of norms to analyze the logical microstructure of arguments. Toulmin's scheme for argumentation and reasoning comprised three main categories: claim, data, and warrant. According to Toulmin, claims are assertions about what exists or what values people hold. The ground for making a claim is derived from data (i.e., facts or observations about the situation under discussion). Warrants explain how the data support the claim using common sense rules, laws, scientific principles, or thoughtfully argued definitions (Hillocks, 2011).

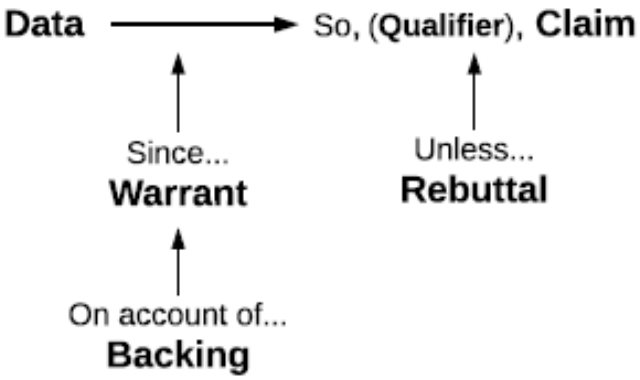

Figure 1: Toulmin's argumentation scheme.

To capture different aspects of human reasoning, Toulmin included three other argument elements: qualifier, rebuttal, and backing (see Figure 1). Qualifiers are linguistic signals (e.g., probably, presumably) that indicate the strength of the relationship between a claim and data conferred by the warrant. Rebuttals denote circumstances in which the general authority of the warrant would have to be set aside. Finally, backing is the knowledge structure from which justifications for the warrants are derived. Without backing, the warrants themselves may not possess either authority or currency.

Toulmin's argument scheme offers an intuitively plausible set of categories and relations for representing the logical structure of arguments and is thought to capture rhetorical structure better than other approaches (Jonassen & Kim, 2009; Ramage & Bean, 1989). It has been commonly used in argumentative writing instruction in the English community (Hillock, 2010; Newell, Beach, Smith, & VanDerHeide, 2011) and has also been widely adopted in research to identify the argumentation elements in students' argumentative essays and/or evaluate the strength of those arguments (Crammond, 1998; Ferretti, MacArthur, & Dowdy, 2000;

McCann, 1989; Nippold & Ward-Lonergan, 2010; Nussbaum & Kardash, 2005, Stapleton & Wu, 2015).

However, cautions have also been raised concerning the reliability of Toulmin's scheme in identifying argumentation elements in texts. For instance, Sampson and Clark (2008) found that some elements in student-generated sample argument can be identified as either a part of the claim, or a qualifier, or even a rebuttal. They also found that when arguments became longer, statements could be classified as new claims or warrants for a preexisting claim. Sampson and Clark thus warned that personal bias in distinguishing among these elements might adversely affect interrater reliability when attempting to classify argumentation elements in texts. Kunnan (2010) reported the classificatory ambiguity of Toulmin's scheme when both explicit and implicit claims are involved in arguments, which may jeopardize rater reliability. Similarly, Simon (2008) noted that implicit claims in argumentation discourse need to be deduced, which may make it difficult to identify relevant data, warrants and backings.

Such reliability issues have prompted researchers to create modified or simplified versions of Toulmin's scheme for argumentation analysis (Nemeth & Kormos, 2001; Nussbaum & Kardash, 2005; Nussbaum & Schraw, 2007; Qin & Karabacak, 2010; Stapleton, 2001; Varghese & Abraham, 1998). For instance, when coding argumentative essays written by undergraduate students at a U.S. university, Nussbaum and Kardash (2005) adopted a modified Toulmin's scheme to accommodate argument structures commonly seen in students' essays wherein an opinion or a conclusion on the main question (final claim) is usually supported by one or more reasons (primary claims) or claims (rebuttals) refuting some potentially opposing opinions to the final claim (counterclaims). They also included supporting reasons or examples which can be used to back up the stated claims. In a similar vein, both Qin and Karabacak (2010) and Stapleton and Wu (2015) used an adapted Toulmin's scheme that comprised six elements: claim, data, counterargument claim, counterargument data, rebuttal claim, rebuttal data. Their schemes highlighted the important roles counterargument claims and rebuttal claims play in defending a claim. In general, these modified versions of Toulmin's scheme, although varied according to the specific focuses in their analyses, have facilitated closer studies of argumentative discourses in educational contexts and provide valuable theoretical and methodological information.

1.2 Rhetorical Structure Theory

Researchers have raised concerns about relying solely on prototypical argumentative discourse schema like those laid out by Toulmin to analyze written arguments because they are generally based on oral argumentation (e.g., Toulmin's schema was developed for legal arguments in the courtroom) and not based on theories of writing or text analysis (Azar, 1999; Brassart, 1996a; 1996b). To address these concerns, researchers have turned to Rhetorical Structure Theory (RST; Mann

& Thompson, 1986; 1987; 1988), which is a theory of text organization commonly used by computational linguists to arrange and connect parts of almost any text type to construct a whole. RST focuses on relationships between discourse elements, including argumentative features, to demonstrate how an entire text functions. Unlike Toulmin's schema, RST can not only define individual discourse functions, but also relationships among these functions. Discourse elements, referred to as text spans in RST, are related through a small set of rhetorical relations. Texts are deconstructed into two types of text spans: the nucleus and the satellite. The nucleus is more text essential than the satellites because satellites can be removed without breaking text coherence, but nuclei cannot (Mann & Thompson, 1988). This generally leads to asymmetrical relationships between nuclei and satellites (i.e., hierarchical relationships). RST analyses require texts to be broken down into text spans and relationships then need to be made between pairs of spans to connect them coherently (Azar, 1999).

An early attempt at exploring the usefulness of RST in persuasive writing was undertaken by Mann and Matthiessen (1991), who noted that the primary goal of a written text functioned as the nucleus and the supplementary material that supported this goal were contained in the satellites. Later, Azar (1999) also confirmed that many RST relations were related to argumentation including evidence, justification, antithesis, and concessions. As a specific example, he argued that under an RST analysis, the claims made in persuasive writing could be treated as nuclei and the arguments used to support them as satellites. A similar approach was adapted by Green (2010) when she modeled argumentative texts with RST. Using hierarchical trees, Green identified how an evidence relationship can link a claim (i.e., the nucleus) with its argument (i.e., the satellites) and a background relationship can link evidence (i.e., the nucleus) with its warrant (i.e., the satellite).

Other research has examined elements of argumentative structure in student essays that also includes graphically depicted structural relationships. For instance, Ferretti et al. (2009) coded structural relationships that were used to differentiate between superordinate and subordinate relations among arguments to distinguish argumentative strategies including argument from example, cause and effect, and consequence. However, Ferretti et al. did not examine the number and types of relationships in reference to human ratings of writing quality.

1.3 Argumentation and Writing Quality

In educational settings, argumentative texts, produced particularly in response to on-demand or impromptu writing tasks, are usually rated by humans using a scoring rubric that attends to either the holistic quality of the texts (holistic scoring) or the quality of specific features or dimensions (analytic scoring). Thus, holistic scores of argumentative writing quality reflect human raters' overall impression of the texts, derived from mentally absorbing and synthesizing all the features therein, i.e., rhetorical, grammatical, and mechanical. Analytic scores, in contrast, measure the

quality of each individual feature or element, isolated to distinguish among a set of subskills (e.g., content development, organization, vocabulary) in argumentative writing (Brown & Abeywickrama, 2010; Wolcott & Legg, 1998).

Annotating discourse elements in argumentative texts has allowed researchers to investigate links between the structural features of argumentative writing and human ratings of writing quality, providing indications about the importance of argumentation in various writing tasks. Previous studies have generally documented links between the number and types of Toulmin elements in student essays and human ratings of writing quality. Cooper et al. (1984), for example, examined a small sample of argumentative papers (N = 10) composed by first-year college students at an American university using Toulmin's scheme. Their findings revealed that less effective essays did not include data and warrants while more effective papers included more elaborative data, warrants, and backings. In a similar study, Crammond (1998) examined 36 argumentative essays written by 6th, 8th, and 10th grade students and found that the number of Toulmin elements (e.g., claims, data, and warrants) increased as a function of grade level. In more recent studies, Nussbaum and Kardash (2005) examined the associations between the inclusion of counterarguments and rebuttals and the ratings of writing quality in first-year college students. Their study revealed that students who employed a greater number of counterarguments and rebuttals in their writing were given higher holistic scores for quality. Similar findings have been reported for second language (L2) writers of English. Qin and Karabacak (2010) analyzed the number and types of Toulmin elements in argumentative essays by undergraduate students (N=130) at a university in China. and reported that the instances of counterarguments and rebuttals showed significant correlations with ratings of essay quality. Likewise, Liu and Stapleton (2014) also documented that the inclusion of counterarguments and rebuttals was positively correlated with the human rated quality of argumentative essays written by a group of college students at a Chinese university.

These studies shed light on the importance of argumentation elements in explaining human ratings of writing quality. However, they tend to examine these argumentative elements in an isolated manner and provide little information on the strength of combining argumentative features, which helps to develop text and argument coherence (Stapleton & Wu, 2015), to predict ratings of essay quality. More importantly, by following a simple categorical schema based on Toulmin annotations, these studies did not consider relationships between argumentative elements as found in rhetorical structure theory approaches.

1.4 Current Study

In the current study, we examine links between human ratings of writing quality and the incidence of argumentative elements in persuasive essays along with relationships among these elements and their distance from one another within an essay. Thus, unlike previous studies, we do not solely examine the incidence of

argumentative elements but also their relationships and distance. Additionally, while we examine links between these elements and holistic score, we also examine links to human ratings related to argumentation strength/organization and introductory elements. Our goal is to better understand how argumentation elements in persuasive essays combine to model human ratings of essay quality. This study is guided by a single research question as follows: Are argumentative elements (their incidence, relationships, and positioning) predictive of human ratings of essay quality?

2. Method

The methods used in this analysis include the initial collection of a corpus of student essays, rating those essays for both holistic and analytic scores, and using a principal component analysis to develop aggregated factors for the analytic scoring items. We follow that by hand annotating the essays for argumentative features (e.g., final claims, primary claims, and data) and the relationships between those essays and then deriving argumentative features to use in a statistical analysis.

2.1 Corpus

The corpus used in the present study comprises 314 essays written by undergraduate, first-year college students (N = 314) at a public research university in the United States. The university was a large state university with a population of mostly White students (~80%). The second largest group of students were African American (~18%) followed by Hispanic students (~2%). Gender breakdown at the university was 49.5% female and 50.5% male. Separate composition classes for non-native speakers of English were required at the university, and these classes were not sampled in this data collection. The students were all native speakers of English who took one of the three composition courses: Composition I, Composition II, and Advanced Composition. Composition I classes were generally taken by first-semester, first-year college students while Composition II classes were taken by second semester students. Advanced Composition classes were available for students who perform well on the written portion of the entrance exam used by the university. Students who pass Advanced Composition were exempted from taking Composition I and II. During data collection, students from these three courses were given 25 minutes to write one persuasive essay on a computer with no outside referencing. Two Scholastic Assessment Test (SAT) writing prompts were used: one prompt was about originality and uniqueness, while the other was about admiring heroes versus celebrities (for complete versions of these two prompts, please see Appendix A). These two prompts were used in retired SAT tests taken by high school students attempting to enter post-secondary institutions in the United States. The prompts were counterbalanced among students such that half of the participants (N = 157) wrote about "originality and uniqueness" while the other half

(N= 157) wrote about "heroes versus celebrities". Upon analyzing the 314 essays, it was obvious that only 298 of the essays used an argumentative structure as expected in a persuasive essay. That is, sixteen of the essays were either too short to contain argumentative structures or the writers did not properly follow the assignment. Thus, we only used the 298 essays that were identified as argumentative for this analysis. Those essays had a mean length of 355.681 words (standard deviation = 117.293).

2.2 Essay Quality Rating

Essays were scored by trained raters on overall quality using a holistic, six-point grading scale, which was a standardized rubric commonly used in assessing SAT essays (see Appendix B) as well as an analytic rubric that looked at sub-components of writing quality (see Appendix C). The holistic scale focused on test-takers' development of a point of view on the issue, critical thinking, use of appropriate examples, accurate and adapt use of language, use of variety of sentence structure, and errors in grammar and mechanics as well as text organization and coherence. The analytic scales included ratings for effective lead, clear purpose, clear plan, topic sentence use, paragraph transitions, organization, unity, perspective, conviction, and grammar and mechanics. Each essay was read by two raters, who were aware of the essay collection methods. The raters had either a master's or a doctoral degree in English and at least two years of experience teaching composition classes at the university-level. All raters were full time faculty within an English Department. For a training purpose, the raters first scored 20 practice essays that were not included in the corpus.

Table 1. Initial inter-rater reliability for analytic and holistic scores (before adjudication)

Item	Exact	Adjacent	Kappa	r
Effective Lead	0.521	0.946	0.656	0.665
Clear Purpose	0.486	0.911	0.657	0.673
Clear Plan	0.476	0.891	0.603	0.623
Topic Sentences	0.498	0.923	0.651	0.666
Paragraph Transitions	0.476	0.879	0.615	0.622
Organization	0.502	0.920	0.639	0.655
Unity	0.473	0.882	0.557	0.573
Perspective	0.502	0.927	0.713	0.719
Conviction	0.534	0.923	0.732	0.739
Grammar, syntax, and mechanics	0.482	0.914	0.703	0.733
Holistic Score	0.594	0.930	0.720	0.735

To measure inter-rater reliability (IRR), a weighted Cohen's Kappa was calculated. After an interrater reliability of Kappa of at least .610 was reached in the training set (described as substantial agreement by Cohen, 1960), the raters scored the essays in the corpus independently. Initial IRR for the majority of dataset after scoring was Kappa < .610 (except *Clear Plan*, see Table 1 for exact and adjacent overlap along with IRR for all items). If score differences between two raters were two points or greater, the raters adjudicated the scores through discussion. If agreement was not reached, the score was not changed. Average scores between the raters for the adjudicated holistic and analytic scores were calculated for each essay and used for the data analysis. There was a small, but significant difference between holistic scores and prompt, t(292.760) = 2.547, $p < .050$, d = .300, such that the originality prompt received lower scores (M = 3.162, SD = .995) than the hero prompt (M = 3.443, SD = .908).

To examine the potential for underlying structures in the ten analytic scores and reduce the analytic scores into a more manageable set of measures, a principal component analysis (PCA) was conducted. Prior to the PCA, correlation analysis was conducted on the ten analytic scores to check for multicollinearity. No scores were found to be highly collinear with each other (absolute r > .899) and thus all analytic scores were included in the PCA. Within the PCA, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy indicated that no variables need to be removed (i.e., all KMO value were above .5) and the overall KMO score = .89 indicated a "meritorious" sample (Kaiser, 1974). The PCA reported a Bartlett's test of sphericity, $\chi 2$ (45) = 1901.408, p < .001, indicating that correlations between the writing fluency indices were sufficiently large for the PCA.

Table 2. PCA loadings for analytic scores

Analytic score	Argument Strength and Organization	Introductory Elements	Grammar and Mechanics
Effective lead	-0,14	**0,96**	0,02
Clear purpose	0,02	**0,91**	0
Clear plan	0,2	**0,73**	-0,02
Topic sentences	**0,62**	0,28	-0,07
Paragraph transitions	**0,94**	-0,22	-0,19
Organization	**0,75**	0,19	0
Unity	**0,67**	0,25	-0,03
Perspective	**0,86**	-0,09	0,17
Conviction	**0,8**	-0,04	0,15
Grammar and mechanics	0,03	0,01	**0,96**

Initial analyses were conducted to ascertain the number of components to extract. Three components had eigenvalues over both Kaiser's and Jolliffe's criterion (1 and .7 respectively) and a visual inspection of the scree plot confirmed the presence of three components. Thus, a three-component solution was sought using an oblique rotation (Promax). The factor loadings after rotation indicated that the first component comprised analytic scores for Unity, Perspective, Conviction, Topic Sentences, Paragraph Transitions, and Organization (labeled Argument Strength and Organization). The second component comprised analytic scores for Effective Lead, Clear Purpose, and Clear Plan (labeled Introductory Elements). The third component consisted of grammar and mechanics score only. These components and the strength of their eigen loadings are presented in Table 2. Because our interest in this paper is strictly on discourse structures, further analyses will only focus on the Argument Strength and Organization component scores and the Introductory Elements scores (both of which were weighted scores based on the eigen scores within the component).

2.3 Essay Annotation Rubric

The 298 essays were annotated by normed raters to gain a comprehensive view of the essays' argumentation structure and assess their argumentation quality. We first developed an annotation rubric to facilitate annotating the essays from three aspects: a) identifying argumentation elements in the essays; b) mapping out the relations among the elements; and c) rating the effectiveness of the elements. Because this study is interested in argumentative structures, we only focus on the argumentation elements in the essays and their relations with one another and we do not analyze the effectiveness of the elements.

Argumentation elements

Our argumentation rubric was adapted from the modified Toulmin models presented in Nussbaum and Kardash (2005) and in Liu and Stapleton (2014). The rubric adopted six elements as the building blocks of the argumentation framework. These were final claim, primary claim, counterclaim, rebuttal, data, and concluding summary. Table 3 presents the definitions and examples for each of these elements. Specifically, we used the term "final claim" in Nussbaum and Kardash (2005) to refer to the student authors' overarching position on the prompt (either "originality and uniqueness " or " heroes versus celebrities"). Visual analyses also indicated that it was a common practice for the students to use multi-level argumentation structures comprising sub-claims of final claims. We labeled these sub-claims "primary claim" as found in Nussbaum and Kardash (2005). Two other types of sub-claims were included: counterclaims and rebuttals. As in Liu and Stapleton (2014), we kept the Toulmin category "data" to denote reasons or examples used to support any claims, be they a final claim, primary claim,

counterclaim or rebuttal. We also included "concluding summary" as an integral argumentation element, given that it was often used in students' essays as a signal to complete an argument. This feature reflected the macrostructure of the argumentative schema structure (Townsend et al., 1993).

Table 3. Definitions and Examples of Argumentation Elements

Elements	Definitions	Examples
Final Claim	An opinion or conclusion on the main question	"In my opinion, every individual has an obligation to think seriously about important matters, even when doing so may be difficult."
Primary Claim	A claim that supports the final claim.	"The next reason why I agree that every individual has an obligation to think seriously about important matters is that this simple task can help each person get ahead in life and be successful."
Counterclaim	A claim that refutes another claim or gives an opposing reason to the final claim.	"Some may argue that obligating every individual to think seriously is not necessary and even annoying as some people may choose to just follow the great thinkers of the nation."
Rebuttal	A claim that refutes a counterclaim.	"Even though people can follow others' steps without thinking seriously in some situations, the ability to think critically for themselves is a very important survival skill."
Data	Ideas or examples that support primary claims, counterclaims, or rebuttals.	"For instance, the presidential debate is currently going on. In order to choose the right candidate, voters need to research all sides of both candidates and think seriously to make a wise decision for the good of the whole nation."
Concluding Summary	A concluding statement that restates the claims	"To sum up, thinking seriously is important in making decisions because each decision has an outcome that affects lives. It is also important because if you think seriously it can help you succeed."

Relationship of argumentation elements

Apart from identifying the argumentation elements in students' essays, our rubric also specified possible relationship among these elements to better delineate the structural features of argumentation. Our approach was inspired by and adapted from Rhetorical Structure Theory (RST) in that the purpose was to examine relationships between discourse elements to demonstrate how an entire text functions beyond simple incidence counts. Two types of relations were defined in our rubric for the argumentation elements in students' essays: elements were either in a hierarchical relation or parallel with each other. Figure 2 diagrams the possible relations for the six elements in our argumentation rubric with parallel (and wide) arrows indicating parallel relations and vertical (and narrow) arrows indicating hierarchical relations. As shown, a primary claim could be used to support the final claim (hierarchical relation), but this primary claim could also be parallel to other primary claims corroborating the same final claim (parallel relation). A counterclaim was defined to be parallel to primary claims given that it is a claim presenting a reason that opposes the final claim but is generally to be refuted by the rebuttal to further defend the final claim. Consequently, a rebuttal cannot function alone in supporting a final claim and was thus considered to form a hierarchical relation with a counterclaim but not directly with the final claim. Data were defined to have a hierarchical relation with all the claims and parallel to each other if they are used to support the same claim.

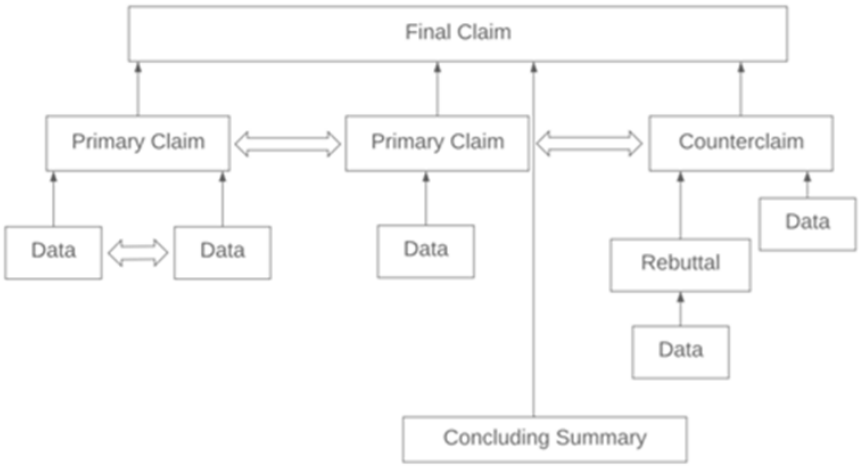

Figure 2: A prototypical diagram for relations among argumentation elements.

2.4 Essay Annotation Procedure

The 298 persuasive essays were coded by two annotators using TagTog (https://www.tagtog.net), a web-based text annotation platform. The two annotators were both undergraduate students majoring in Applied Linguistics. Before annotation, the two annotators were given a 2-hour training on the annotation rubric and Tagtog. Tagtog allowed the raters to highlight sections of a text and annotate that text with discourse element tag (e.g., final claim). Tagtog also allowed raters to indicate hierarchical and parallel relations among the annotated discourse elements. Norming sessions were developed wherein the annotators were asked to annotate sample persuasive essays (36 in total) on their own before they met with an expert annotator to compare their annotation results and discuss the annotations where disagreement arose. Once normed, the annotators worked independently and coded the essays in opposite orders to avoid recency effects.

To measure inter-rater reliability, weighted Cohen's Kappa coefficients were calculated between the two raters for the categories of final claim, primary claim, data, concluding summary, and non-annotated. Counterclaims and rebuttals were rare in the data and were combined into the category of claim. When calculated agreement between raters, we took into account adjacent overlap among annotations. For instance, rater A may code the string of words 'a b c d e f g h i j k' as a final claim and 'l m n o p q' as a claim. In contrast, rater B may code the string of words 'a b c d e f g h' as a final claim and the string of words 'i j k l m n o p g' as claim. Because there is strong overlap between the two annotations, the raters are likely in agreement although not perfect agreement. Thus, for this work, we coded 66% overlap between two annotations as agreement.

The weighted Cohen's kappa for all the annotations was (k) =.643, $p < .001$, while unweighted Cohen's kappa was (k) = .615, $p < .001$, indicating fair to good strength of agreement between the annotators (Fleiss e al., 2003). To calculate inter-rater agreement within each category we used the formula found in Ferretti et al. (2009, agreement = agreements/ agreements + disagreement within one instance). Percentage agreements for each category are presented in Table 4.

Table 4. Agreement scores for each category (percent)

Category	Agreement
Final claim	0.69
Primary claim	0.60
Data	0.79
Concluding summary	0.72
Non-annotated	0.83

Disagreements between the two annotators were adjudicated using a third, expert annotator. The expert annotator was a PhD student with years of experience teaching and researching writing structures and quality. In the case of a disagreement, the expert annotator made the final decision as to which annotation was correct. The adjudicated annotations were used in the analyses of the data.

2.5 Deriving Argumentation Features from Annotations

One crucial issue in the current study was to quantify the argumentative structures and their relationships. To quantify the number of argumentative structures, we identified the incidence of each argumentative structure in each essay as identified by the raters. We also calculated the number of discourse segments in the text that were not annotated as discourse elements (i.e., non-annotated segments). We used raw counts per essay for this feature. To calculate the relationship among the argumentation elements, we calculated the number of hierarchical and parallel relations each element shared with other elements as annotated by the expert raters. We used raw counts per essay for these features. We were also interested in the distance between argumentation elements that had relationships with each other because this distance would give us a better understanding of how closely or loosely the related elements were knitted together. We presumed that if discourse elements had greater hierarchical distance between them, this would indicate more supporting information was presented to support the relationship. This feature was operationalized by first locating the locations of two linked elements in the text and then calculating the number of characters between the two elements. To illustrate, consider the text location of a final claim of {55, 94} where 55 represents the beginning position of the final claim in references to the number of characters since the beginning of the essay and 94 stands for its ending position. The text distance between this final claim and a supporting primary claim with a location of {152, 170} would be 58 characters (i.e., 152 - 94 = 58). Using this approach, we attempted to calculate distance scores between elements that shared hierarchical or parallel relations in argumentation. In practice, we found that this approach only worked for final claims, because almost all essays had final claims making it possible to calculate distances to subordinate categories such as primary claims or data. This was not the case with other categories, which were more infrequent in the data. As an example, if an essay did not include a primary claim, it was impossible to calculate hierarchical distance for that category for that essay. Instead of removing essays for which we could not calculate distance, we only calculated distance for final claims. Specifically, we averaged the distance scores for each subordinate category under final claims.

2.6 Statistical Analysis

Our goal was to examine if discourse features and relationships among those features were predictive of holistic score and the Argument Strength/Organization

and Introductory Elements component scores. To do so, we used a Random Forest Regression Model. Random forests are powerful and flexible ensemble classifiers based on decision trees (Breiman, 2001). We selected random forest models over more traditional statistical methods like linear regression models because random forest models do not require the data to be normally distributed. Much of the data coded for this analysis was not normally distributed because all essays did not contain all the potential discourse features. For instance, essays may not include a concluding summary or even primary claims.

Random forest algorithms produce binary trees using different bootstrap samples of the data. Each node in the tree is split using the best among a subset of predictors randomly chosen at that node. The response variable (in this case the holistic or the component scores) in a random forest regression is modeled using an ensemble of regressions. These regressions create rules for divisions of the observation until the predictions have a minimum amount of node impurity. Random forest models also report on variable importance using a mean decrease in GINI impurity importance metric. GINI impurity is based on the weighted mean of how the splitting criterion used for each variable can lead to an individual trees' prediction improvement. Random forests have two parameters: numbers of tree and number of variables.

Prior to running random forest models, we selected blocks of variables based on theoretical considerations. These included four blocks: 1) variables related to counts, 2) variables related to hierarchical relationships, 3) variables related to parallel relationships and, 4) variables related to hierarchical distance (see Table 5 for blocks and variables). We used the CARET package (Kuhn et al., 2016) in R (R Core Team, 2017) to develop random forest regression models and incrementally added each block to the model in the order above. Model training and evaluation were performed using k-fold, cross-validation techniques. Specifically, we used k-fold cross validation (10-fold CV), which works well on small datasets like the one reported here. K-fold cross-validation estimates performance on predictions for data not used in the training of the model. The "k" in k-fold controls the number of subsets into which data is split. Each subset is used once as a test set while the other subsets are combined to form the training set. Thus, in k-fold cross-validation, k number of models are fitted and evaluated allowing for k estimates of performance.

In 10-fold CV, 10 models are developed and evaluated for each fold providing 10 estimates of a model's performance on the ten different test sets. These estimates are reported using summary statistics including root mean squared error (RMSE) and mean absolute error (MAE) between the observed and modeled holistic and analytic scores. R-squared (R2) is also reported and can be used examine the amount of variance explained by the developed model. We set our number of trees to the default (N =500). We set our number of variables to include all possible variable combinations for each data set. After running each random forest regression model per block, we removed variables from the block that did not

improve the model and then added in the next block of variables. We used the function rsamples() from CARET to compare the RMSE and R2 reported for each block in succession and the compare_models() function to quantitatively test model performance.

Table 5. Blocks and variables used in analyses

Block	Variable
1	Final claim: Count
	Primary claim: Count
	Counterclaim: Count
	Rebuttal: Count
	Data: Count
	Concluding summary: Count
	Non-annotated: Count
2	Final claim: Hierarchical count
	Primary claim: Hierarchical count
	Counterclaim: Hierarchical count
	Rebuttal claim: Hierarchical count
3	Primary claim: Parallel count
	Counterclaim: Parallel count
	Data: Parallel count
4	Final claim: Hierarchical distance

In 10-fold CV, 10 models are developed and evaluated for each fold providing 10 estimates of a model's performance on the ten different test sets. These estimates are reported using summary statistics including root mean squared error (RMSE) and mean absolute error (MAE) between the observed and modeled holistic and analytic scores. R-squared (R2) is also reported and can be used examine the amount of variance explained by the developed model. We set our number of trees to the default (N =500). We set our number of variables to include all possible variable combinations for each data set. After running each random forest regression model per block, we removed variables from the block that did not improve the model and then added in the next block of variables. We used the function rsamples() from CARET to compare the RMSE and R2 reported for each block in succession and the compare_models() function to quantitatively test model performance.

3. Results

Descriptive statistics (Mean and Standard Deviation) are reported in Table 5 for all the variables in this analysis. As might be expected, each essay had, on average, around one final claim, two primary claims, and a concluding summary (although the mean for concluding summary was below 1). The greatest number of counts were for data. As noted in the methods, counterclaims and rebuttals were infrequent. Final claims had the greatest number of subordinate elements followed by primary claims and primary claims had the greatest number of parallel relations (i.e., other primary claims).

Table 6. Descriptive stats for all variables

Variable	Mean	SD
Holistic score	3.304	0.961
Introductory elements	0.000	1.000
Argument strength and introduction	0.000	1.000
Final claim: Count	1.003	0.058
Primary claim: Count	1.765	1.217
Counterclaim: Count	0.060	0.265
Rebuttal: Count	0.067	0.300
Data: Count	2.477	0.996
Concluding summary: Count	0.822	0.383
Final claim: Hierarchical count	3.158	0.931
Primary claim: Hierarchical count	1.889	1.399
Counterclaim: Hierarchical count	0.077	0.354
Rebuttal claim: Hierarchical count	0.050	0.248
Primary claim: Parallel count	1.772	1.245
Counterclaim: Parallel count	0.070	0.374
Data: Parallel count	0.691	1.318
Final claim: Hierarchical distance	1782.155	668.258

3.1 Correlations

We provide a correlation matrix for the variables used in this analysis and the human scores of writing quality of interest (i.e., holistic score, introductory elements component score, and argument strength and organization). The matrix (see Table 7) shows that many variables demonstrated at least a small effect size with the human scores and few variables were highly multi-collinear.

Table 7. Correlational matric for all variables in analysis

Variable	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Holistic Score	0,716	0,860	-0,018	0,113	0,040	0,040	0,188	0,193	0,096	0,213	0,147	0,035	0,056	0,135	0,053	0,082	0,213
2. Introductory Elements Component Score		0,707	-0,056	0,118	0,036	0,034	0,124	0,010	0,109	0,071	0,150	0,039	0,043	0,132	0,082	0,023	0,047
3. Argument Strength and Organization			-0,014	0,129	0,022	0,022	0,193	0,323	0,053	0,309	0,143	0,013	0,048	0,149	0,052	0,083	0,254
4. Final claim count				0,059	-0,013	-0,013	0,031	0,027	0,085	0,053	0,046	-0,013	-0,012	0,057	-0,011	-0,030	-0,019
5. Primary claim count					-0,133	-0,132	0,510	-0,047	-0,047	0,520	0,901	-0,138	-0,128	0,985	-0,045	-0,310	0,405
6. Counterclaim count						0,965	0,005	-0,026	0,106	0,016	-0,136	0,954	0,926	-0,050	0,873	-0,024	0,028
7. Rebuttal count							-0,006	-0,013	0,059	0,022	-0,135	0,935	0,907	-0,049	0,829	-0,033	0,027
8. Data count								-0,121	0,074	0,663	0,696	0,000	0,012	0,512	0,045	0,572	0,325
9. Concluding summary count									-0,168	0,400	-0,106	-0,023	-0,047	-0,029	0,017	-0,082	0,118
10. Nonannotated count										-0,071	0,003	0,078	0,094	-0,030	0,128	0,100	-0,054
11. Final claim hierarchical count											0,419	-0,006	0,009	0,542	0,094	0,180	0,366
12. Primary claim hierarchical count												-0,139	-0,139	0,882	-0,062	0,001	0,396
13. Counterclaim hierarchical count													0,838	-0,067	0,798	-0,028	0,027
14. Rebuttal hierarchical count														-0,050	0,797	-0,024	0,029
15. Primary claim parallel count															0,071	-0,314	0,393
16. Counterclaim parallel count																-0,058	0,031
17. Data parallel count																	0,033
18. Final claim hierarchical distance normed																	

3.2 Holistic scores

When the first block (count variables) was entered into a random forest regression model to predict holistic score, the number of variables per regression tree that performed the best in the optimization procedure for the training set was two: data count and primary claim count (see Table 8 for model performance). Increasing the number of variables decreased model performance. The two predictive variables from Block 1 were included in a subsequent random forest regression model that also included the variables from Block 2 (hierarchical count variables).

Table 8. Model performance metrics by block: Holistic score

Blocks	Variables	Variable importance	RMSE	R- squared	MAE
1	Data: Count	15.69			
	Primary claim: Count	15.27	0.951	0.073	0.799
1:2	Final claim: Hierarchical count	23.47			
	Primary claim: Hierarchical count	17.5	0.943	0.075	0.787
2:3*	Final claim: Hierarchical count	23.47			
	Primary claim: Hierarchical count	17.5	0.943	0.075	0.787
2:3, 4	Final claim: Hierarchical distance	89.31			
	Final claim: Hierarchical count	24.21	0.975	0.064	0.801

*best performing model

The number of variables per regression tree that performed the best in the optimization procedure for the training set for this model was two: hierarchical count for final claims and hierarchical count for primary claims (see Table 8 for model performance). Increasing the number of variables decreased model performance. There was no significant difference between the model for variables from Block 1 and the model that included variables from Block 1 and 2, $t(9) = .278$, $p = .788$. However, since the MSE was lower, the R2 was higher, and the MAE was lower for the model with Block 2 variables, those variables were included with the Block 3 variables (parallel count variables) in the next random forest regression model and the variables from Block 1 were removed.

The random forest regression for Blocks 2 through 3 reported the same output as the model from Blocks 2 (i.e., no parallel count variables were included in the new model, see Table 8). Thus, the two variables from the Block 2 model were added to the variable from Block 4 (hierarchical distance for final claim) in a final random forest regression model.

3.3 Introductory elements

When the first block (count variables) was entered into a random forest regression model to predict the introductory elements score, the number of variables per regression tree that performed the best in the optimization procedure for the training set was two: data count and primary claim count (see Table 9 for model performance). Increasing the number of variables decreased model performance. The two predictive variables from Block 1 were included in a subsequent random forest regression model that included the variables from Block 2 (hierarchical count variables).

The number of variables per regression tree that performed the best in the optimization procedure for the training set was two: data count and hierarchical count for primary claims (see Table 9 for model performance). Increasing the number of variables decreased model performance. There was no significant difference between the model for variables from Block 1 and the model that included variables from Block 1 and 2, $t(9) = .233$, $p = .821$. However, since the RMSE was lower, the R2 was higher, and the MAE was lower for the model with Block 1 variables, those variables were included with the Block 3 variables (parallel count variables) and the Block 2 variables were removed.

The number of variables per regression tree that performed the best in the optimization procedure for the training set for this model was two: primary claim parallel count and data count (see Table 9 for model performance). Increasing the number of variables decreased model performance. There was no significant difference between the model for variables from Block 1 and the model that included variables from Block 1 and 3, $t(9) = .258$, $p = .802$. However, since the RMSE was lower, the R2 was higher, and the MAE was lower for the model with Block 1 and 3 variables, those variables were included with the Block 4 variable (hierarchical distance for final claim) in the final model.

The number of variables per regression tree that performed the best in the optimization procedure for the training set for the final model was two: hierarchical distance for final claims and primary claim parallel count (see Table 9 for model performance). Increasing the number of variables decreased model performance. There was no significant difference between the model for variables from Block 1 and 3 and the model that included variables from Block 3 and 4, $t(9) = .401$, $p = .698$. However, since the RMSE and the MAE scores were similar between the two models, but the R2 was higher in the model with Block 3 and 4 variables, this model

was considered the best model to predict introductory element scores. In total, the model from Blocks 3 and 4 explained around 10 percent of the variance (r = .312).

Table 9. Model performance metrics by block: Introductory elements score

Blocks	Variables	Variable importance	RMSE	R-squared	MAE
1	Data: Count	12.14			
	Primary claim: Count	12.08	0.985	0.04	0.818
1:2	Primary claim: Hierarchical count	15.35			
	Data: Count	11.26	0.993	0.033	0.825
1, 3	Primary claim: Parallel count	14.14			
	Data: Count	11.73	0.984	0.059	0.818
1, 3:4*	Final claim: Hierarchical distance	83.48			
	Primary claim: Parallel count	22.38	1.001	0.097	0.827

*best performing model

3.4 Argument Strength and Organization

When the first block (count variables) was entered into a random forest regression model to predict argument strength and organization score, the number of variables per regression tree that performed the best in the optimization procedure for the training set was two: concluding summary count and data count (see Table 10 for model performance). Increasing the number of variables decreased model performance. The two predictive variables from Block 1 were included in a subsequent random forest regression model that included the variables from Block 2 (hierarchical count variables).

The number of variables per regression tree that performed the best in the optimization procedure for the training set for the second model was two: concluding summary count and hierarchical count for final claims (see Table 10 for model performance). Increasing the number of variables decreased model performance. There was no significant difference between the model for variables from Block 1 and the model that included variables from Block 1 and 2, $t(9) = .002$, $p = .998$. However, since the R2 was higher for the model with Block 1 and 2 variables, those variables were included with the Block 3 variables (parallel count variables).

The random forest regression for Blocks 1 through 3 reported the same output as the model from Blocks 1 and 2 (i.e., no parallel count variables were included in the model, see Table 10). Thus, the two variables from the Block 1 and 2 models were added to the variable from Block 4 (Hierarchical distance for final claim) in a final random forest regression model.

The number of variables per regression tree that performed the best in the optimization procedure for the training set for the final was two: hierarchical distance for final claims and hierarchical count for final claims (see Table 10 for model performance). Increasing the number of variables decreased model performance. There was no significant difference between the model for variables from Block 1, 2, and 4, and the model that included variables from Block 1, 2 and 3, $t(9) = .978$, $p = .354$. However, since the RMSE was lower, the R2 was higher, and the MAE was lower for the model with Block 1 and 2 variables, this model was considered the best model to predict argument strength and organization scores. In total, the model from Blocks 1 and 2 explained around 16 percent of the variance ($r = .399$).

Table 10. Model performance metrics by block: Argument strength and organization score

Blocks	Variables	Variable importance	RMSE	R-squared	MAE
1	Concluding summary: Count	30.16			
	Data: Count	16.32	0.928	0.154	0.772
1:2*	Concluding summary: Count	23.33			
	Final claim: Hierarchical count	22.19	0.928	0.159	0.772
1:3	Concluding summary: Count	23.33			
	Final claim: Hierarchical count	22.19	0.928	0.159	0.772
1:2, 4	Final claim: Hierarchical distance	92.8			
	Final claim: Hierarchical count	29.28	0.971	0.095	0.802

*best performing model

4. Discussion

The goal of this study was to investigate the importance of argumentation features in predicting various aspects of human ratings of essay quality. In similar fashion to previous studies, we examined the incidence of argumentation features using a modified Toulmin schema. Like some previous studies, we also examined relationships between argumentation features (Ferretti et al., 2009), although, unlike previous studies, we investigated their associations with essay quality scores. Lastly, unlike previous studies, we examined distance between related features and their prediction of writing quality scores. Our analyses indicated that, in general, the presence or absence of argumentation features was not as important in

predicting aspects of essay quality scores when compared to relationships among features and the distance between those relationships.

Our initial analysis examined links between overall scores for essay quality and argumentation features and their relationships. We expected, based on previous studies (Cooper et al., 1984; Crammond, 1998; Qin & Karabacak, 2010), that the incidence of argumentation features would significantly predict human ratings of essay quality. Our correlation analysis supported this to a degree with two incidence scores showing significant correlations (counts of concluding summaries and data, see Table 7). However, the strongest correlations were yielded for variables that calculated relationships between argumentation features. Specifically, hierarchical counts and hierarchical distances for final claims showed the strongest correlations. The hierarchical count for final claims along with the hierarchical count for primary claims were the strongest predictors for modeling overall ratings of essay quality as demonstrated by the variable importance metrics derived from the random forest models. These two variables, combined, explained around 8% of the total variance in the human scores for essay quality. The results indicate that essays greater hierarchical argumentative structures for both final and primary claims are predicted to have higher holistic scores.

Our introductory elements score comprised analytic scores for effective lead, clear purpose, and clear plan. Three argumentation features demonstrated at least weak effect sizes with the introductory element component (see Table 7). These variables demonstrated that a strong introduction was linked to primary claim counts, data counts, non-annotated counts, hierarchical counts for primary claims, and parallel counts for primary claims. The hierarchical distance for final claims along with the parallel count for primary claims were the strongest predictors for modeling introductory elements component scores ratings. These two variables, combined, explained around 10% of the total variance in these scores. The results indicate that stronger introductory elements in essays are predicted by the distance between final claims and their relations and the number of parallel primary claims in the essay.

Our final analysis looked at links between argumentation features and argument/organization component scores. The correlational analysis (see Table 7) indicated that two variables demonstrated at least a moderate relationship with argumentation/organization scores and five other variables were significantly, although weakly, correlated with argumentation/organization scores. These variables indicated that higher argumentation/organization scores were the result of greater number of primary claims, data elements, and concluding summary. In addition, higher scores were association with a greater number of hierarchical relations for final claims and primary claims and greater parallel relationships for primary claims and hierarchical distance for final claims. Our statistical model demonstrated that two of these variables (*Concluding summary: Count* and *Final claim: Hierarchical count*) explained 16% of the variance in the argumentation/

organization scores. The model indicated that greater counts for concluding summary and greater hierarchical counts for final claims were strong predictors of argument strength and organization scores.

Overall, these findings indicate that the strongest predictors of writing quality scores (holistic, introductions, and argumentative/organization scores) were operationalizations of relationships between argumentation features and not the incidence of argumentation features. Our relationship variables were based on three constructs: hierarchical relationships, parallel relationships, and distance between related discourse structures. Of these three, the strongest indicators of writing quality scores seemed to be the hierarchical count variables, which showed positive correlations with writing scores. The findings indicate that argument structures that have a greater number of hierarchical relations (i.e., deeper argument structures) are related to essay quality scores. Strong relationships were also reported for parallel count features, showing that a greater number of parallel relationships (for instance, more parallel relations between primary claims) is predictive of higher essay scores. Our feature related to hierarchical distance for final claims was also a significant predictor, indicating that the greater the textual distance (measured by character length) between a subordinate discourse element (e.g., a primary claim) and its superordinate discourse element (e.g., a final claim), the greater the writing quality score. Intuitively, this makes sense because greater distance would indicate more supporting data between the end of a superordinate discourse element and the beginning of its related subordinate element. A worry, however, may be that this operationalization of distance is unknowingly measuring text length in general, which is a strongly related to writing quality scores (Authors). To examine this possibility, we conducted post-hoc correlations between text length and distance variables and found only weak correlations ($r < .30$) with text length, limiting the effect of text length on these variables.

Beyond our operationalization of relationships between argumentation features, another strength of this study is the multivariate nature of the modeling used in the analysis. As noted, most previous studies examined how counts of single argumentation elements were related to writing quality scores. In this analysis, we used multiple variables to predicting writing quality scores and found that, for the most part, simple count variables were not predictive when relationship variables were included (the exception being concluding summary counts). It should be noted, however, that our models explained a relatively small amount of the variance for most of our writing quality scores (8% of holistic writing quality, 10% of introductory elements, and 16% of argumentative/organization scores). While not the focus of this study, future research should consider multivariate analyses that include textual elements beyond discourse elements including text cohesion features, syntactic complexity, grammar and mechanics accuracy, and lexical sophistication. It is likely that these features explain much of the variance not captured by our discourse element features. In addition, we were not able to

control for individual differences in our data (e.g., working memory, vocabulary knowledge, inferencing skills), which may also explain variance in writing and argumentation skills. Lastly, we did not examine the effectiveness ratings for the argumentative elements. A post-hoc analysis showed that these ratings reported similar correlations to the argumentative structure (i.e., small to medium effect sizes), and we presume that adding them into fuller models of writing quality would increase the amount of variance explained.

Our study has a number of practical implications that go beyond better understanding the interaction of argumentation features and human ratings of essay quality. Foremost are pedagogical applications for the teaching of writing. While many students are instructed on the development and use of argumentation features, the findings from this study indicate that pedagogy may be enhanced by focusing on the development of relationships between features and tracking of distance between related features. Another potential application for the findings from this study relates to automatic essay scoring and writing evaluation. The argumentation features reported here were hand annotated. However, it is possible that the incidence of argumentation features and their relationships could be automatically annotated using natural language processing (NLP) techniques and these automatic annotations could be incorporated into educational technologies to provide writers with automated feedback about the argumentation features in their essays (see Song, Deane, & Beigman-Klebanov, 2017). Such additions are needed because most automated writing feedback writing tools focus on lower-level textual features to both assess writing quality scores and provide feedback (Strobl et al., 2019).

There are also limitations to this study. For instance, the essays used in this analysis were the product of a timed writing task in which students were only allowed 25 minutes to write an essay. It is possible that 25 minutes is not enough time to produce deeper features of argumentation. For instance, counterarguments and rebuttals require time and effort to embed into the larger structure of an essay (Shehab & Nussbaum, 2015). Likewise, collecting data from more advanced writers (e.g., graduate writers) may afford for deeper argumentation structures to annotated. Limitations also pop up in the scoring procedures used. Specifically, each essay was only scored or annotated by two raters, which may not provide enough variance to fully explain argumentative quality and elements. Additionally, the raters depended on rubrics, which may be less reliable that other methods of evaluation like comparative judgments (Van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2016). Relatedly, Kappa values from the rubrics prior to adjudication were on the low side indicating the potential for noisy data, which may lead to problems with generalizability.

5. Conclusion

This study has provided a greater understanding of how argumentation features related to claims, evidence, and concluding summaries are associated with various aspects of writing quality. Unlike previous studies, our focus was not solely on predicting writing quality based on the incidence of argumentation features but also on the relationships between these features. We find that, in most cases, it is not the presence of argumentation features that is predictive of writing quality but rather the relationships between superordinate, subordinate, and parallel features and the distances between them. This finding has not only theoretical value but also practical value in terms of pedagogical approaches and automated writing feedback.

While the findings help to understand how argumentation features interact with essay quality, there are still many follow up studies that need to be undertaken to help support and replicate the findings reported here. Primarily, these studies should focus on replicating the current findings with a larger sample size. While ~300 essays provide a strong foundation for analysis, thousands of essays will need to be annotated, coded, and analyzed. It is not just purely the number of essays needed for replication but also the types of writing tasks. The current study focuses solely on independent essay writing, which, while a common writing task, is but one of many writing tasks that require the use of argumentative elements. Future studies should consider how the annotation scheme used in this study can be applied to source-based writing and other similar tasks. Beyond sample size and tasks, future studies should also examine how the methods used in this study generalize to populations other than first-year college composition students and to a greater number of prompts. Lastly, while this study examined the production of argumentation features and the relationships between the features, it did not consider whether the elements were effective or not. Future studies should investigate differences between effective and ineffective argumentative elements on judgments of writing quality.

6. References

Applebee, A. N.(1984). Writing and reasoning. *Review of Educational Research Winter*, 54(4), 577-596. https://doi.org/10.3102/00346543054004577

Azar, M. (1999). Argumentative text as rhetorical structure: An application of rhetorical structure theory. *Argumentation*, *13*(1), 97-114.

Brassart, D. G. (1996). Didactique de l'argumentation écrite: Approches psycho-cognitives. *Argumentation*, *10*(1), 69-87. https://doi.org/10.1007/bf00126160

Brassart, D. G. (1996). Does a prototypical argumentative schema exist? Text recall in 8 to 13 years olds. *Argumentation*, *10*(2), 163-174. https://doi.org/10.1007/bf00180723

Berrill, D. P. (1996). Reframing argument from the metaphor of war. In P. Berrill (Ed.), *Perspectives on written argument* (pp. 171–187). Creskill, NJ: Hampton Press.

Breiman, L. (2001). Decision-tree forests. *Machine Learning*, *45*(1), 5-32.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (Vol. 10). White Plains, NY: Pearson Education.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational Psychology Measurement, 20*, 37–46. https://doi.org/10.1177/001316446002000104

Cooper, C.R., Cherry, R., Copley, B., Fleischer, S., Pollard, R., Sartisky, M. (1984). Studying the writing abilities of a university freshman class: strategies from a case study. In Beach, R.,Bridwell, L.S. (Eds.), *New Directions in Composition Research*. The Guilford Press, New York, pp. 19-52. https://doi.org/10.1017/s0142716400006263

Crammond, J. (1998). The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication*, 15, 230-268. https://doi.org/10.1177/0741088398015002004

Crasnich, S., & Lumbelli, L. (2005). The reflection-response in enhancing argumentation ability. *L1-educational Studies in Language and Literature*, *5*(2), 147-174. https://doi.org/10.1007/s10674-005-0918-5

Currie, P. (1996). Fullness and sound reasoning: Argument and evaluation in a university content course. In P. Berrill (Ed.), *Perspectives on written argument* (pp. 121-137). Creskill, NJ: Hampton Press.

Ferretti, R. P., Lewis, W. E., & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' argumentative writing strategies?. Journal of Educational Psychology, 101(3), 577. https://doi.org/10.1037/a0014702

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology*, *92*(4), 694. https://doi.org/10.1037/0022-0663.92.4.694

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions Wiley-Interscience. *Hoboken, NJ*. https://doi.org/10.1002/0471445428

Gárate, M., & Melero, A. (2005). Teaching How to Write Argumentative Texts at Primary School. *Studies In Writing Effective Learning and Teaching of Writing*, 323–337. https://doi.org/10.1007/978-1-4020-2739-0_22

Green, N. L. (2010). Representation of argumentation in text with rhetorical structure theory. *Argumentation*, *24*(2), 181-196. https://doi.org/10.1007/s10503-009-9169-4

Hillocks, G. (2010). " EJ" in Focus: Teaching Argument for Critical Thinking and Writing: An Introduction. *The English Journal*, *99*(6), 24-32.

Hillocks, G. J. (2011). *Teaching argumentative writing, Grades 6-12: Supporting claims with relevant evidence and clear reasoning.* Portsmouth: Heinemann.

Hirvela, A. (2017). Argumentation & second language writing: Are we missing the boat? *Journal of Second Language Writing*, *36*, 69–74. https://doi.org/10.1016/j.jslw.2017.05.002

Hyland, K. (1997). *A genre description of the argumentative essays*. RELC Journal 21 (1): 66–78. http://dx.doi.org/10.1177/003368829002100105

Jonassen, D. H., & Kim, B. (2010). Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development*, *58*(4), 439-457. https://doi.org/10.1007/s11423-009-9143-8

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., ... & Kenkel, B. the R Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C (2016) Caret: classification and regression training. *R package version*, 6-0.

Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, 27, 183–189. https://doi.org/10.1177/0265532209349468

Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing*, *33*, 21–34. https://doi.org/10.1016/j.jslw.2016.06.004

Liu, F., & Stapleton, P. (2014). Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. *System, 45*, 117-128. https://doi.org/10.1016/j.system.2014.05.005

Mann, W. C., & Matthiessen, C. M. (1991). Functions of language in two frameworks. *Word*, *42*(3), 231-249. https://doi.org/10.1080/00437956.1991.11435839

Mann, W. C., & Thompson, S. A. (1986). Relational propositions in discourse. *Discourse processes*, *9*(1), 57-90. https://doi.org/10.1080/01638538609544632

Mann, W. C., & Thompson, S. A. (1987). *Rhetorical structure theory: A theory of text organization* (pp. 87-190). Los Angeles: University of Southern California, Information Sciences Institute.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243-281. https://doi.org/10.1515/text.1.1988.8.3.243

McCann, T. M. (1989). Student argumentative writing knowledge and ability at three grade levels. *Research in the Teaching of English*, 62-76.

Nemeth, N., & Kormos, J. (2001). Pragmatic aspects of task-performance: the case of argumentation. *Language Teaching Research*, 5, 213-240. https://doi.org/10.1177/136216880100500303

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.

Nippold, M. A., & Ward-Lonergan, J. M. (2010). Argumentative writing in pre-adolescents: The role of verbal reasoning. *Child Language Teaching and Therapy*, 26, 238–248. https://doi.org/10.1177/0265659009349979

Nussbaum, E. M., & Kardash, C. M. (2005). The Effects of Goal Instructions and Text on the Generation of Counterarguments During Writing. *Journal of Educational Psychology*, *97*(2), 157–169. https://doi.org/10.1037/0022-0663.97.2.157

Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *The Journal of Experimental Education*, *76*(1), 59-92. https://doi.org/10.3200/jexe.76.1.59-92

Pessoa, S., Mitchell, T. D., & Miller, R. T. (2017). Emergent arguments: A functional approach to analyzing student challenges with the argument genre. *Journal of Second Language Writing*, *38*, 42–55. https://doi.org/10.1016/j.jslw.2017.10.013

Qin, J., & Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System*, *38*(3), 444–456. https://doi.org/10.1016/j.system.2010.06.012

R Core Team (2017). R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria*.

Ramage, J. D., & Bean, J. C. (1989). *Writing arguments: A rhetoric with readings*. New York: Macmillan.

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, *92*(3), 447–472. https://doi.org/10.1002/sce.20276

Shehab, H.M. & Nussbaum, E.M. (2015). Cognitive load of critical thinking strategies. *Learning and Instruction 35*, 51-61. https://doi.org/10.1016/j.learninstruc.2014.09.004

Simon, S. (2008). Using Toulmin's Argument Pattern in the evaluation of argumentation in school science. *International Journal of Research & Method in Education*, 31, 277-289. https://doi.org/10.1080/17437270802417176

Song, Y., Deane, P., & Beigman Klebanov, B. (2017). Toward the automated scoring of written arguments: Developing an innovative approach for annotation. *ETS Research Report Series, 2017(1)*, 1-15. https://doi.org/10.1002/ets2.12138

Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students: Insights about assumptions and content familiarity. *Written communication*, *18*(4), 506-548. https://doi.org/10.1177/0741088301018004004

Stapleton, P., & Wu, Y. A. (2015). Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes*, *17*, 12–23.

https://doi.org/10.1016/j.jeap.2014.11.006

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33-48. https://doi.org/10.1016/j.compedu.2018.12.005

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge: Cambridge university press.

Townsend, M. A., Hicks, L., Thompson, J. D., Wilton, K. M., Tuck, B. F., & Moore, D. W. (1993). Effects of introductions and conclusions in assessment of student essays. *Journal of Educational Psychology*, *85*(4), 670. https://doi.org/10.1037/0022-0663.85.4.670

Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*. https://doi.org/10.1080/0969594x.2016.1253542

Varghese, S. A., & Abraham, S. A. (1998). Undergraduates arguing a case. *Journal of Second Language Writing*, *7*, 287-306. https://doi.org/10.1016/s1060-3743(98)90018-2

Walton, D., Reed, C., and Macagno, F. (2008), Argumentation Schemes, Cambridge: Cambridge University Press.

Wolcott, W., & Legg, S. M. (1998). *An Overview of Writing Assessment: Theory, Research, and Practice*. Urbana, IL: National Council of Teachers of English.

Appendix A

Prompt One:

We value uniqueness and originality, but it seems that everywhere we turn, we are surrounded by ideas and things that are copies or even copies of copies. Writers, artists, and musicians seek new ideas for paintings, books, songs, and movies, but many sadly realize, "It's been done." The same is true for scientists, scholars, and business people. Everyone wants to create something new, but at best we can hope only to repeat or imitate what has already been done. Can people ever be truly original? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

Prompt Two:

Having many admirers is one way to become a celebrity, but it is not the way to become a hero. Heroes are self-made. Yet in our daily lives we see no difference between "celebrities" and "heroes." For this reason, we deprive ourselves of real role models. We should admire heroes—people who are famous because they are great—but not celebrities—people who simply seem great because they are famous. Should we admire heroes but not celebrities? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

Appendix B

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.

Appendix C: Analytical rating form

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

We present here a description of the grade as a guide using the example of *does not meet the set criterion in any way* versus *meets the set criterion in every way*. For example, a grade of 1 would relate to not meeting the criterion in any way, and a grade of 4 would relate to somewhat meeting the criterion. The distance between each grade (e.g., 1-2, 3-4, 4-5) should be considered equal. Thus, a grade of 5 (*meets the criterion*) is as far above a grade of 4 (*somewhat meets the criterion*) as a grade of 2 (*does not meet the criterion*) is above a grade of 1 (*does not meet the criterion in any way*).

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

Part	Score
1. Introduction	
1.1 Effective Lead The introduction begins with a surprising statistic, a provocative quotation, a vivid description, an engaging fragment of dialogue, or some other device to grab the reader's attention and point toward the thesis.	1 2 3 4 5 6
1.2 Clear Purpose The introduction includes one or two sentences that provide essential background information and establish the significance of the discussion.	1 2 3 4 5 6
1.3 Clear plan *The introduction ends with a thesis statement that provides a claim about the topic and a preview of the support and organizational principle to be presented in the body of the essay.*	1 2 3 4 5 6
2. Body	
2.1 Topic Sentences	1 2 3 4 5 6

Each paragraph includes a sentence (often at the beginning) that connects with the thesis and makes a comment on one of the points outlined in the introduction.	
2.2 Paragraph transitions Each topic sentence is preceded by a phrase, clause, or sentence that links the current paragraph with the previous one, stressing the relationship between the two.	1 2 3 4 5 6
2.3 Organization The body paragraphs follow the plan set up in the introduction, underscoring the organizational principle.	1 2 3 4 5 6
2.4 Unity The details presented throughout the body support the thesis and do not stray from the main idea.	1 2 3 4 5 6
3. Conclusion	
3.1 Perspective The writer summarizes the key points that collectively sustain the thesis and stress its significance.	1 2 3 4 5 6
3.2 Conviction The author re-establishes the significance of the discussion as it pertains to the thesis.	1 2 3 4 5 6
4. Correctness	
4.1 Grammar, syntax, and mechanics The writer employs correct Standard American English, avoiding errors in grammar, syntax, and mechanics.	1 2 3 4 5 6