frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 19 June 2025
DOI 10.3389/feduc.2025.1569529

OPEN ACCESS

EDITED BY
Maria Cutumisu,
McGill University, Canada

REVIEWED BY
Mehrdad Yousefpoori-Naeim,
University of Alberta, Canada
Man-Wai Chu,
University of Calgary, Canada

*CORRESPONDENCE
Amy Burkhardt
amy.burkhardt@cambiumassessment.com

RECEIVED 01 February 2025
ACCEPTED 15 May 2025
PUBLISHED 19 June 2025

CITATION
Burkhardt A, Han S, Woolf S, Boykin A, Rijmen F and Lottridge S (2025) Standards-aligned annotations reveal organizational patterns in argumentative essays at scale.
Front. Educ. 10:1569529.
doi: 10.3389/feduc.2025.1569529

Standards-aligned annotations reveal organizational patterns in argumentative essays at scale

Amy Burkhardt*, Suhwa Han, Sherri Woolf, Allison Boykin, Frank Rijmen and Susan Lottridge

Cambium Assessment, Inc., Washington, DC, United States

While scoring rubrics are widely used to evaluate student writing, they often fail to provide actionable feedback. Delivering such feedback—especially in an automated, scalable manner—requires the standardized detection of finer-grained information within a student's essay. Achieving this level of detail demands the same rigor in development and training as creating a high-quality rubric. To this end, we describe the development of annotation guidelines aligned with state standards for detecting these elements, outline the annotator training process, and report strong inter-rater agreement results from a large-scale annotation effort involving nearly 20,000 essays. To further validate this approach, we connect annotations to broader patterns in student writing using Latent Class Analysis (LCA). Through this analysis, we identify distinct writing patterns from these fine-grained annotations and demonstrate their meaningful associations with overall rubric scores. Our findings show promise for how fine-grained analysis of argumentative essays can support students, at scale, in becoming more effective argumentative essay writers.

KEYWORDS

argumentative writing assessment, writing feedback, rubrics, assessment, machine learning, automated scoring and feedback

1 Introduction

Scoring rubrics are the cornerstone of standardized writing evaluation, with millions of U.S. middle-school students' argumentative essays scored according to established criteria, such as those detailed in the Smarter Balanced Assessment Consortium (2022) used in 10 U.S. states. Yet, rubrics are not always self-explanatory (Andrade, 2005). While rubrics effectively standardize scoring, their broad categories often mask the diverse ways students can achieve a particular score level. This inherent limitation means that rubric scores alone provide insufficient guidance for improvement–students may know their current performance level but remain uncertain about specific steps for improvement. For example, effective feedback as conceptualized by Hattie and Timperley (2007), address three essential questions that students may ask: what are the goals (*Where am I going?*), what progress has been made (*How am I going?*), and what activities are needed to improve progress (*Where to next?*). While rubrics may reflect learning goals (Brookhart and Chen, 2015), they fall short in addressing the latter two questions. These two questions rely on the identification and analysis of fine-grained writing patterns.

To address these gaps, we propose a parallel system of guidelines for identifying specific elements of argumentative writing–referred to as annotation guidelines. While rubrics offer a broad overview of writing quality, annotations allow for a more detailed analysis of specific components, providing actionable feedback that helps students improve their writing strategies and identify areas for development. While existing

corpora have already been labeled according to various argumentative theories (e.g., Stab and Gurevych, 2014; Crossley et al., 2022), this paper introduces annotation guidelines specifically designed to align with widely used rubrics. This tailored approach bridges the gap between theoretical frameworks and practical classroom application, making it possible to link fine-grained annotations to standards-aligned feedback.

The annotations based on these guidelines have broad applications, including use by practitioners, researchers, and even machine learning models. Annotated collections of essays can inform the training or fine-tuning of models, enabling the automatic identification of argumentative components and the generation of meaningful feedback. Recent studies show that such annotations are well-suited for machine learning applications, with models demonstrating strong performance in predicting labels (e.g., Ormerod et al., 2023).

After a brief overview of related work, this paper presents a detailed example of the annotation process, followed by three main investigations: (1) outlining a replicable method for developing annotation guidelines, including training procedures and inter-annotator agreement results, (2) analyzing whether these annotations can uncover meaningful patterns in student writing on a large scale, and (3) whether these writing patterns exhibit meaningful associations with overall rubric scores. Together, these investigations aim to establish and validate a robust, scalable method for analyzing student writing.

1.1 Related works

A foundational contribution to argumentation annotation comes from Stab and Gurevych (2014, 2017), who introduced one of the first corpus of annotated persuasive essays, and it has subsequently been used by other researchers for argumentative modeling tasks (e.g., Nguyen and Litman, 2018). Their annotation scheme identifies three key argumentative components: major claim, claim, and premise. The corpus consists of 322 essays written by students who shared their work on an online forum seeking support to improve their argumentative writing. Inter-rater agreement was evaluated at the sentence level, which was described as an approximation, as argument components may not align neatly with sentence boundaries, and individual sentences can contain multiple components. Exact agreement was highest for major claims (97.9%), followed by premises (91.6%) and claims (88.9%). These values declined when accounting for chance. The authors also reported Fleiss' κ (Fleiss, 1971) and Krippendorff's α (Krippendorff, 2004): Fleiss' κ was 0.87 for major claim, 0.83 for premise, and 0.64 for claim; Krippendorff's α was 0.81 for major claim, 0.82 for premise, and 0.52 for claim.

More recently, Crossley et al. (2022) developed an annotation framework tailored to student essays, based on Toulmin's argumentative framework (Toulmin, 2003). Their corpus, the PERSUADE corpus, contains over 25,000 essays from grades 6–12 across 12 different prompts. The sample includes essays from all score points on the rubric. Their guidelines includes the following discourse elements: lead, position, claim, counterclaim, rebuttal, evidence, and concluding statement. Due to segmentation differences between raters, aligning annotations to evaluate for inter-rater reliability proved challenging. To calculate agreement, the authors defined inter-rater reliability (IRR) as instances where there was at least a 50% overlap between the discourse elements annotated by the first and second raters. IRR scores ranged from 0.68 for claims, counterclaims, and rebuttals, to 0.73 for leads and evidence, 0.74 for concluding statements, and 0.81 for positions. When raters evaluated the effectiveness of each discourse element, agreement was lower, with weighted kappa values ranging from 0.17 for claims to 0.43 for evidence.

To date, no argumentation annotation scheme has fully captured the broader structure of argumentative essays as represented in widely used U.S. K–12 rubrics -- both in terms of the range of argumentative elements and their alignment with the criteria used in statewide summative assessments. This study aims to address that gap.

1.2 An example of annotations in an argumentative essay

To provide readers with a clear understanding of how annotations appear within an essay, we first present a visual demonstration to establish context and help illustrate their application in practice before introducing the detailed criteria necessary for reliable labeling.

The following essay demonstrates the application and interpretation of the annotations within an argumentative essay written in defense of the use of computers.[1] In this example all seven elements of argumentative writing are represented (Figure 1): *Introduction*, *Controlling Idea*, *Evidence*, *Elaboration*, *Opposing Position*, *Transitions*, and *Conclusion*. In the first paragraph, the student introduces the controlling idea (♦, yellow), and then contextualizes the argument with introductory text by outlining their plan for their argument (★, green). The second paragraph begins with a topic sentence that restates a point from the introduction. This type of sentence functions as a transition (♥, light green), helping to guide the reader through the essay. Notably, such transitions may also appear at the end of a paragraph to signal what is coming next. The transition is followed by elaboration (♣, pink). Within this paragraph, the student also includes a piece of evidence (▮, blue), and then further synthesizes this evidence with more elaboration (♣, pink). The third paragraph contains a similar pattern, beginning with a transition (♥, light green), followed by elaboration (♣, pink) and evidence (▮, blue). The fourth paragraph reflects a pattern very similar to the second paragraph. Lastly, the first annotation of the final paragraph indicates an opposing position (▶, purple). This is followed by a rebuttal, in the form of conclusion text (✚, red).

Figure 1 provides an example of how a student approaches their argumentative essay. This visual representation can vary for each student's essay, highlighting finer-grained information that can be used to provide personalized feedback, especially when key

1 This essay was retrieved from the training materials of the Automated Student Assessment Prize (ASAP) competition in 2012. The contents of which can be downloaded here: https://www.kaggle.com/c/asap-aes/data.

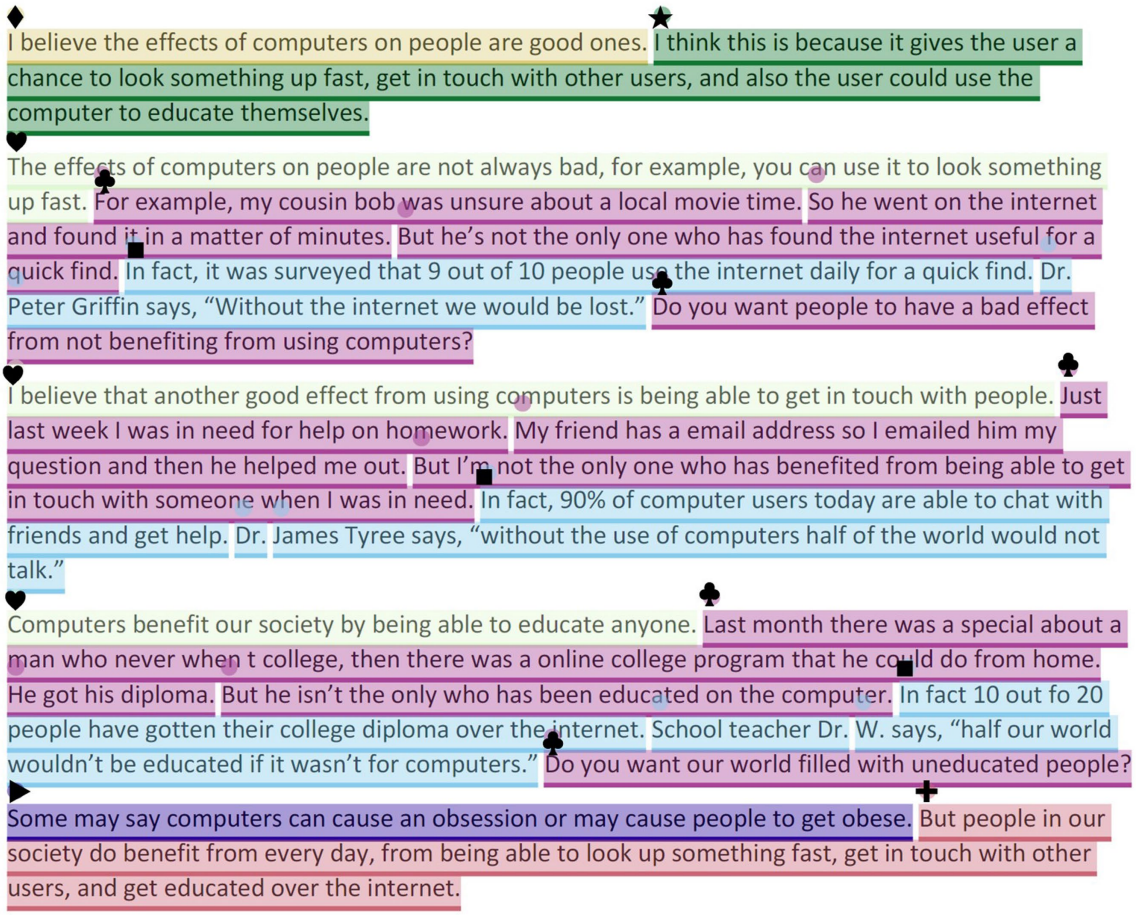

FIGURE 1
Example essay.

argumentative elements are missing. The next section outlines the development of the standards-aligned annotation guidelines and describes their application to a sample of approximately 20,000 essays.

2 Methods and materials

2.1 The development of annotation guidelines

As mentioned in the previous section, the seven elements of argumentative writing were obtained from criteria delineated from two widely used rubrics: Integrating College and Career Readiness (ICCR; presented to in Appendix A), as well as the Smarter Balanced Argumentative Rubric (Smarter Balanced Assessment Consortium, n.d.). We selected these two rubrics as they have been widely implemented for interim and summative use in over 15 states. The selection of these two rubrics is also justifiable as these two rubrics showed high continuity across the criteria within each rubric level and they both incorporate the same key elements of argumentative writing. Integration of the Smarter Balanced rubric into these guidelines was particularly crucial in the development for two reasons: First, integration provides alignment and application to millions of students in the United States. Second, this integration afforded us access to supporting materials to inform the details of the guidelines. Of note, the focus was on the argumentative writing rubric from 6th–8th grade for ICCR and 6th–11th grade for Smarter Balanced. While other states may not use the ICCR or Smarter Balanced rubrics, their writing rubrics do focus on similar elements in writing (e.g., Florida Department of Education, 2025; Texas Education Agency, 2022).

The ICCR rubric includes three dimensions: (1) Purpose, Focus, and Organization; (2) Evidence and Elaboration; and (3) Conventions of Standard English. The first two dimensions are relevant to identifying key compositional elements that comprise the annotation guidelines. Within the Purpose, Focus, and Organization dimension, relevant elements include the main claim, acknowledgement of the opposing position, the use of transitional strategies to clarify relationships between ideas, and the presence of an introduction and conclusion. In the Evidence and Elaboration dimension, key features include the use of evidence and elaborative techniques to support the claim and demonstrate understanding of the source material. These elements appear at every performance level of the rubric, but the expectations for their quality become more rigorous at higher levels. Additional aspects of writing–such as style, tone, and precise language–are addressed in the rubric but are not captured in the discrete annotation elements.

In developing the guidelines, the following practical requirements were considered: they should be straightforward to minimize the cognitive load of the annotators; they should only require background knowledge in essay-scoring and not require additional linguistic expertise; they should not be time intensive; and they should be applicable to under-developed essays, where the student is still developing their ideas and organizational structure. In alignment with these practical considerations and supported by research suggesting that ratings of quality tend to be low in such contexts (Crossley et al., 2022), the quality of these argumentative elements are not assessed. Figure 2 below summarizes the guidelines that were developed in this study.

The first column in the figure presents the seven annotation tags, each referring to one of the seven elements of argumentative writing: *Introduction*, *Controlling Idea*, *Evidence*, *Elaboration*, *Opposing Position*, *Conclusion*, and *Transitions*. These seven elements were empirically derived from a qualitative review of the two rubrics alongside materials used to train professional hand-scorers for these rubrics. The next column, "Definition" provides a brief description of each annotation. These definitions were informed by large-scale hand-coring practices and further refined through synthesis of the rubric language and supporting documentation. The final two columns of the guidelines -"What should be highlighted" and "What should not be highlighted." - were developed iteratively, incorporating lessons learned from initial annotation trials and insights from related Smarter Balanced materials. For instance, the criteria highlighting *Elaboration*–such as relevant commentary, definitions of key terms, rhetorical questions, or rebuttals of an opposing position – draw from the Smarter Balanced elaboration guidelines (Smarter Balanced Assessment Consortium, 2022). These two columns provide concrete criteria to help guide annotators through the more difficult and challenging decisions.

As we refined the annotation guidelines, two additional rules were developed. The first relates to the unit of annotation. Specifically, each annotation tag should be applied at the sentence level, allowing for efficiency in the task and limits the cognitive load in isolating phrases that belong to a certain annotation tag. Another rule emerged from this first decision: Assign only a single tag to each sentence. While this second rule reduces the cognitive burden of a rater in applying several tags to a single sentence, it does pose a problem when there is more than a single argumentative element present in a sentence.

To address potential annotation inconsistencies that can be caused by multiple elements in a sentence, we established the following hierarchy for assigning argumentation tags: *Controlling Idea*, *Opposing Position*, *Evidence*, *Elaboration*, and *Transition*. The order within the hierarchy was designed to prioritize the most rhetorically central components of argumentative writing. *Controlling Idea* and *Opposing Position* were emphasized as foundational to the overall argument. *Evidence* was placed above *Elaboration* because it requires additional cognitive effort, such as reviewing and selecting source material. *Transitions* were ranked lowest, as they introduced no new content and instead served a connective function.

Introduction and *Conclusion* were excluded from this hierarchy. These elements were treated as separate structural components of the essay. Including them in the tagging hierarchy would have introduced potential confusion, particularly because they sometimes contain embedded argumentative components. For example, a piece of *Evidence* may appear in the *Introduction*

Argumentation Element	Definition	What should be highlighted	What should not be highlighted
★ Introduction	Context to explain the issue	• Plan for an argument, such as listing out subtopics (organizational outline) • Use of rhetorical devices to help reader feel like this issue is important • Attention grabbing devices (which may include citation/quotation of sources) • Stage setting sentences	The controlling idea in the introduction is tagged as a controlling idea (see below)
♦ Controlling Idea	Overarching statement about the debatable issue	• The stand that the author is taking on an issue or topic • Taking center of the road position • Flip flops (highlight both claims)	All of the sub-topics of the controlling idea (this is either elaboration or introduction)
♣ Elaboration	Explanations, elaborations, and interpretations of evidence and controlling idea	• General supports of claim • Reasons • Sub-claims • Rhetorical questions or other rhetorical devices (e.g., allusions, appeal to logos, repetition, parallelism, fragments for effect) to enhance arguments • Relevant commentary and examples • Definitions for related terms • Rebuttal of opposing position in support of controlling idea	Extraneous information, such as an entire paragraph that does not seem relevant to the controlling idea or a sentence that jumps out as being off-topic.
■ Evidence	Integration of evidence, including data, statistics or similar study results, and quotations to support the controlling idea or sub-topics	• Citation/quotation of sources • Data from sources • Paraphrasing authors as a citation.	
▶ Opposing Position	Acknowledge the opposing side of the argument	• Acknowledgement of the opposing position (not in intro or conclusion) • Sentences that explore or explain the opposing position	
✚ Conclusion	Summation the controlling idea	• Summary of sub-topics, evidence, and elaboration • Tug-at-heartstring moments or call to action • Use of rhetorical devices to garner support (which may include citation/quotation of sources)	Introduction of new ideas
♥ Transitions	Signposts to help guide the reader through the development of the essay	• Sentences at the end of a paragraph signaling what is coming next • Sentences at the beginning of the paragraph reiterating a sub-claim detailed in the introduction • Sentences with no new information intending to create coherence or structure • Sentences that re-state introductory organizational phrases to guide the sub ideas of each paragraph	Sentences that use transitional clauses but contain evidence, elaboration, or opposing positions. These should be tagged by using the hierarchy listed at the top of the table
None of the Above	Off-topic and extraneous information		

FIGURE 2
Annotation guidelines.

TABLE 1 Count and length statistics for essays within each prompt and grade level.

Prompt ID	Grade	Number of responses	Mean sentences (SD)	Mean paragraphs (SD)
A	Grade 6	1,925	18 (10)	4 (3)
B	Grade 6	1,931	19 (10)	5 (3)
C	Grade 6	1,926	18 (10)	4 (3)
D	Grade 7	1,948	20 (10)	4 (2)
E	Grade 7	1,953	21 (11)	4 (2)
F	Grade 7	1,937	20 (11)	4 (2)
G	Grade 8	1,929	22 (11)	4 (2)
H	Grade 8	1,953	22 (11)	4 (3)
I	Grade 8	1,949	20 (10)	5 (2)

as an attention-grabber, or in the *Conclusion* as a final persuasive appeal.

2.2 Annotation training materials and process

To support the annotation process, additional training materials were developed alongside the guidelines, all tailored to the 6–8 grade band. Applying a unified training approach across this grade band is consistent with the design of the ICCR rubric, which was explicitly developed for grades 6 through 8. While it is expected that students' writing will show increasing sophistication as they progress through grades, this developmental progression is not encoded through separate rubrics for each grade.[2] Instead, calibration occurs through the examples used in training; providing raters with grade-specific exemplars that ground expectations at each performance level. For the annotation process, we follow the same logic to use consistent guidelines across the grade band 6-8.To create training materials for the annotation work, hand-scoring training materials previously used to train raters in accordance with the original rubric were repurposed. Based on the repurposed hand-scoring training and qualification materials, the researchers compiled an anchor set and practice sets consisting of 10 essays each across prompts across all three grade levels. Prior to the training session, the developers of the annotation guidelines independently annotated all training papers, after which they reviewed and discussed the annotations to arrive at a final consensus annotation for each essay.

The training session with 13 experienced hand-scoring professionals, referred to as "annotators," occurred over a two-day period. One of the authors of the annotation guidelines, who possesses extensive expertise in training hand-scoring is referred to herein as the "trainer." The trainer conducted an in-depth review of the contents of the annotation guidelines and then presented a series of five anchor essays, each accompanied by a detailed description of the annotations for each sentence. Next, the annotators provided their own annotations for the first practice set, which were subsequently reviewed as a group against the "true" annotations to clarify any discrepancies. This process was repeated for three additional practice sets. Upon completion of the fourth practice set, the trainer assessed that all annotators had demonstrated sufficient comprehension of the task, and instructed them to complete the fifth practice set, and to then commence the annotation process for the essays assigned to them. Following the training session, the annotators annotated the essays using the INCEpTION annotation software (Klie et al., 2018), which was hosted in a secure Amazon Web Service environment.

2.3 Data for annotation

The study utilized essay responses written to nine prompts used in a statewide summative assessment program in a Southern state of the United States across three academic years (2018–2019, 2020–2021, 2021–2022). The writing portion of the assessment typically includes two to three source passages, and students are asked to construct an argument that integrates evidence from both sources. This format is consistent across the prompts used in the program. The writing prompt itself is part of a larger assessment system that also includes components in reading, mathematics, and science.

Stratified random sampling was used on the sum of the three rubric dimensions to ensure representation of all summed score points because the higher score points were rare in the student population. Very short essays, and essays that were flagged as non-attempts were also removed. As a result, a total of 17,451 essays (approximately 2,000 for each of nine prompts) were annotated. Fifteen percent of these annotated essays were randomly assigned to obtain a second rating, which was later used to compute a human rater agreement. Table 1 presents the number of essays as well as the average number of sentences and paragraphs for responses per prompt. The nine prompts were distributed across grade 6, grade 7, and grade 8.

Table 2 presents the score point distributions for two dimensions of the rubric used in this project: The dimension of *Purpose, Focus, and Organization* and the dimension of *Evidence and Elaboration*. As shown in the table, the distribution of scores are similar across the two dimensions, in that very few responses received four points, and that most essays received two points across all nine prompts.

2 It should be noted that one grade-level distinction reflected in both the ICCR rubric and the Common Core State Standards is that grade 6 students are not required to address opposing positions (Council of Chief State School Officers & National Governors Association Center for Best Practices, 2010, p. 42).

TABLE 2 Score point distribution (%) for each prompt and rubric dimension (n = 17,451 essays).

Prompt ID	Grade	Organization				Evidence and Elaboration				Conventions		
		1	2	3	4	1	2	3	4	0	1	2
A	6	38	51	10	1	65	31	4	0	8	36	56
B	6	46	45	8	1	62	32	5	0	7	31	62
C	6	44	46	10	0	58	37	5	0	5	47	48
D	7	16	80	4	0	35	64	2	0	3	24	73
E	7	35	60	5	0	56	41	3	0	3	27	70
F	7	36	59	4	0	54	43	3	0	3	42	55
G	8	26	50	22	2	28	54	17	2	6	21	73
H	8	23	59	17	2	28	59	12	1	5	28	67
I	8	30	53	16	1	37	51	12	0	6	21	74

2.4 Analysis of annotator experiment

Because there is currently no industry standard for evaluating sentence-level agreement in annotation tasks within educational writing contexts – as evidenced by the variety of methods employed by Stab and Gurevych (2014, 2017) as well as Crossley et al. (2022)–we extended this line of inquiry by exploring three different approaches to assess interrater agreement using Cohen's kappa (Cohen, 1960). Kappa, more specifically quadratic-weighted kappa, is regarded as a standard measure of agreement in automated essay scoring (e.g., Williamson et al., 2012), making it a strong candidate for adaptation to sentence-level annotations. However, how to best apply this metric at the sentence level remains an open question. By implementing three different computational approaches, our aim was to examine the extent of variation in interrater agreement that might result from each method. A series of analyses was conducted to evaluate the accuracy and consistency of the annotations on the sampled essays. First, three different agreement statistics were computed to evaluate patterns in the distribution of the agreement of the sentence-level annotations. The first two statistics rely on Cohen's kappa, while the third examines the agreement rate of all labels in an essay. Cohen's kappa in the current context requires two pieces of information: The proportion of sentences where the two raters agreed with one another, and the expected agreement of the sentences based on chance (calculated based on the independent probabilities of the ratings by the two annotators).

For calculating Cohen's kappa, contingency tables were used to obtain the distribution of ratings between two annotators. The rows and columns can be referred to as the rating variables, and herein lies the difference between the two approaches. The first approach (Approach #1) considered each of the seven annotation labels as rating variables; all sentences from all essays were included in the same 7×7 contingency table. The second approach (Approach #2) computes a 2×2 contingency table for each of the seven annotation labels. For each sentence, the presence of an annotation label was marked by a value of '1' and the absence is marked by a value of '0'. In this second approach, a single, aggregated kappa value is computed by averaging all the kappa values together. Both aggregated and disaggregated values are presented in this paper. Finally, the third approach (Approach #3) was computed for each essay in adherence to the following: For each sentence, if the annotation label matches for the two annotators, mark this agreement as a '1' and otherwise, mark the sentence as a disagreement, '0'. Then, to arrive at an agreement rate for each essay, take the average of all values. A single agreement statistic is computed by averaging across all essays. The first two approaches aggregate the annotations independent of the essay itself, while the last approach examines annotation agreement within each essay and aggregate across the essays. To guide the interpretation of results, we refer to commonly cited conventions that suggest the following benchmarks: values between 0.21–0.40 represent fair agreement, values between 0.41–0.60 represent moderate agreement, 0.61–0.80 substantial agreement, and values above 0.80 near-perfect agreement (Landis and Koch, 1977).

2.5 Analysis of annotations

Annotated sentences offer valuable insights into the organizational structure and elaboration strategies within individual essays. At the aggregate level, these annotations can be leveraged to identify and differentiate patterns in student writing across a broader population. In this study, we use latent class analysis (LCA) to uncover distinct writing patterns that not only highlight meaningful variations within the population but also reveal how specific writing behaviors align with rubric criteria. The ability to connect fine-grained annotations to broader trends strengthens the validity and credibility of the annotation guidelines. This connection offers evidence that the annotation guidelines are capturing meaningful aspects of student writing.

LCA is a statistical method to analyze multivariate categorical data (Lazarsfeld, 1968; Bishop et al., 2007). It explains the statistical dependencies between categorical indicators by assuming that the population can be partitioned into a set of mutually exclusive homogeneous subgroups or classes. Each class is characterized by a set of response probabilities for each of the indicator variables. The pattern of (conditional) probabilities are the basis for interpreting each of the underlying subgroups. Specifically, applied to annotations (or more precisely, features based on raw annotations, see section 2.5.1), they can provide insight into how annotations cluster together and are indicative of distinct writing approaches. Individual students can be assigned to one of the classes based on

their posterior probabilities to belong to each of the classes given their annotation pattern. The LCA results can also be investigated in comparison to existing scoring criteria to ensure the validity of the proposed annotation framework.

2.5.1 Feature engineering

The raw format annotation data are a sequence of nominal categorical variables. While this sequence data format provides details about where each annotation appears in the essays, as well as its neighboring annotations (i.e., which annotations it is next to), such information must be converted into a format suitable for LCA. In this study, we adopted a rule-based feature engineering approach that automatically converts the sequence of annotations into multivariate categorical features. The rules were developed after qualitative investigation of the essays, contextualizing features in terms of the location and relationships with other annotations.

As the initial step for the feature derivation, we drew on the notion of introduction, body and conclusion paragraphs to describe the location of annotations, as these are used as critical elements in the rubric. Specifically, we defined a paragraph as an introduction if the combined proportion of Introduction and Controlling-Idea tags was larger than 0.6. Similarly, a conclusion paragraph was assigned to any paragraph whose proportion of Conclusion tag exceed 0.6. Paragraphs that were classified as neither introduction nor conclusion were categorized as body paragraphs.

This paragraph classification then served as the foundational variables for our feature set; for each paragraph type (i.e., introduction, body, or conclusion), we created a feature and established mutually exclusive categorical levels to reflect the essay's attribute relevant to the corresponding paragraph. For example, an essay received a level of "intro_introduction_no_controlling_yes_other" on its "introduction" feature if the essay's introduction paragraph consists of *Introduction*, with no *Controlling Idea*. In addition to the three features, we established an additional feature that describes the characteristic of an essay's *Controlling-Idea* in relation to the paragraph variables (e.g., "The essay has *Controlling Idea* in a body paragraph"). Appendix B describes these categorical features in more detail.

2.5.2 Latent class analysis

In this study, LCA was employed to the above-mentioned feature set to discern distinct subgroups, or classes, of argumentative essays. LCA identifies latent classes that are characterized by a set of response probabilities for each of the indicator variables. The prior probabilities and class conditional probabilities are typically estimated during the Expectation–Maximization (EM) optimization process (Linzer and Lewis, 2011). Once these model parameters are estimated, students can be assigned to a latent class based on their posterior class membership probabilities. The posterior probability is defined as a function of prior probabilities of latent class membership and the likelihood of an observation given class conditional probabilities of categorical indicators. Let $i\left(i=1,2,\ldots,N\right)$ denote individual observation and $c\left(c=1,2,\ldots,C\right)$ denote latent class. C can be determined *a priori*, or can be determined empirically by evaluating model fit for a sequence of latent class models with increasing C. In practice, model fit, stability and interpretability are taken into account when determining C, similar to how the number of factors are determined in exploratory factor analysis.

A vector of observed responses for individual i is denoted as R_i. LCA describes the posterior probability that observation i belongs to class c, given the observed responses R_i, $\hat{P}\left(c/R_i\right)$, as the following:

$$\hat{P}\left(c/R_i\right)=\frac{\hat{P}\left(c\right)f\left(R_i;\hat{\pi}_c\right)}{\sum_{l=1}^{C}\hat{P}\left(l\right)f\left(R_i;\hat{\pi}_l\right)}=\frac{\hat{P}\left(c\right)f\left(R_i;\hat{\pi}_c\right)}{P\left(R_i\right)}$$

where $\hat{P}\left(c\right)$ is the prior probability of belonging to class c, $f\left(R_i;\hat{\pi}_c\right)$ is the likelihood having the set of responses R_i given class conditional probability estimates $\hat{\pi}_c$, and $P\left(R_i\right)$ is the marginal probability of the observed responses.

2.5.3 Model fitting and parameter estimation

Given that LCA requires pre-specification of the number of latent classes, we fit a range of LCA models, from four to ten latent classes to the feature set aggregated across nine prompts. This study used the R package poLCA (version 1.6.0.1; Linzer and Lewis, 2011) to fit the models as the package can handle polytomous categorical variables. To avoid converging to local optima during EM estimation, this study conducted 30 replications for each model. The model with the largest likelihood was chosen as the final model for the corresponding class.

2.5.4 Score-conditional posterior probability of latent class membership

The relationship between the latent classes and essay scores is further delineated by computing posterior probabilities of latent classes given essay scores. The score-conditional posterior probability using Bayes Theorem is calculated as the following:

$$\hat{P}(c/s)=\frac{\hat{P}\left(c\right)P\left(s|c\right)}{P\left(s\right)}$$

where $\hat{P}\left(c\right)$ is the estimated class probability for latent class c, $P\left(s|c\right)$ is the conditional probability of observing essay score s given latent class c, and $P\left(s\right)$ is the marginal probability of the essay score s.

TABLE 3 Annotator accuracy across five training samples (Approach # 1).

Annotator	Training sample				
	1	2	3	4	5
1	0.69	0.80	0.76	0.79	0.82
2	0.65	0.62	0.78	0.73	0.79
3	0.62	0.69*	0.61	0.72	0.88
4	0.52	0.79	0.89	0.96	0.96
5	0.67	0.80	0.80	0.67	0.88
6	0.64	0.63*	0.64	0.62	0.68
7	0.78	0.75	0.91	0.77	0.78
8	0.72	0.84	0.85	0.78	0.78
9	0.79	0.88	0.85	0.82	0.79
10	0.69	0.80	0.80	0.82	0.83
11	0.76	0.77	0.83	0.86	0.85
12	0.50	0.86	0.79	0.84	0.82
13	0.70	0.72	0.78	0.90	0.64*

An asterisk (*) indicates the annotator did not complete all essays.

3 Results

3.1 Annotator accuracy and consistency

To provide evidence that annotators accurately applied labels to the sentences according to the annotation guidelines, we report agreement statistics for the five practice sets used in the training phase. Specifically, we utilized the INCEpTION software to compute Cohen's kappa between the labels assigned by each annotator for each sentence and the "true" labels, as determined by the consensus of the researchers who developed the annotation guidelines (Approach #1).

Table 3 presents Cohen's kappa results for each practice set. Our findings indicate that for the first practice set, annotators exhibited a moderate (0.5) to a substantial (0.78) level of agreement with the consensus labels. Notably, we observed an overall increase in agreement for annotators as they progressed through subsequent practice sets. By the fourth and fifth practice sets, all annotators demonstrated an increased level of agreement with the consensus labels, with kappa values ranging from substantial (0.78) to near-perfect agreement (0.96).

Regarding annotator consistency, Table 4 presents results from the three different approaches, based on the 15% sample of essays that were labeled by two raters. On average, Approach #3 reflected the highest agreement, with values ranging from 0.75 to 0.79, indicating substantial agreement. Even though Approach #2 resulted in the lowest agreement values of the three approaches, ranging from 0.60 to 0.70, these values are nonetheless moderate to substantial. This suggests that even the lower-bound estimate of annotator consistency supports that annotators were able to consistently apply the argumentation labels to sentences within an essay. Across these three approaches, Prompt I exhibited the highest agreement, and the grade 6 prompts (A, B, and C) exhibited the lowest agreement.

Of the three approaches used to compute inter-rater reliability, Approach #2 yielded the most conservative (i.e., lower bound) estimate of agreement. This approach was selected for detailed reporting, as it provides a cautious interpretation of annotator consistency. Table 5 presents Cohen's kappa statistics for each of the annotation labels, across all prompts. Generally, annotators exhibited high agreement for the labels of *Introduction*, *Conclusion*, and *Controlling Idea*.

TABLE 4 Annotator consistency.

Prompt ID	Grade	Kappa 1[a]	Kappa 2[b]	Kappa 3[c]
A	6	0.62	0.55	0.75
B	6	0.60	0.59	0.72
C	6	0.60	0.52	0.73
D	7	0.70	0.68	0.78
E	7	0.70	0.67	0.78
F	7	0.66	0.63	0.76
G	8	0.67	0.63	0.76
H	8	0.69	0.67	0.78
I	8	0.71	0.68	0.79
Average		0.66	0.62	0.76

[a]Cohen's kappa statistic computed from a single 7 × 7 contingency table; [b]Cohen's kappa statistic computed from seven 2 × 2 contingency tables, averaged across all annotation labels; [c]Exact agreement rate averaged across all essays.

TABLE 5 Kappa by argumentative tag (Approach #2).

Prompt ID	Grade	Control. idea	Intro.	Elab.	Evid.	Opp. position	Trans.	Conc.
A	6	0.67	0.70	0.59	0.64	0.27	0.25	0.76
B	6	0.72	0.78	0.52	0.48	0.27	0.51	0.86
C	6	0.59	0.68	0.58	0.58	0.09	0.32	0.77
D	7	0.73	0.77	0.67	0.65	0.60	0.51	0.86
E	7	0.79	0.82	0.64	0.66	0.56	0.42	0.83
F	7	0.78	0.71	0.61	0.67	0.47	0.37	0.81
G	8	0.74	0.81	0.61	0.60	0.44	0.34	0.85
H	8	0.78	0.76	0.65	0.67	0.61	0.42	0.78
I	8	0.80	0.75	0.66	0.69	0.47	0.50	0.87

Cohen's Kappa computed according to Approach # 2. Control. Ida = Controlling Idea; Intro. = Introduction; Elab. = Elaboration; Evid. = Evidence; Opp. Position = Opposing Position; Trans. = Transition; Conc. = Conclusion.

TABLE 6 Contingency table of annotation labels.

	Control. idea	Intro.	Elab.	Evid.	Opp. position	Trans.	Conc.	No tag
Controlling Idea	**1,901**	275	195	30	5	2	107	3
Introduction	306	**4,868**	775	183	55	7	1	4
Elaboration	176	784	**18,680**	1,786	943	289	547	29
Evidence	29	143	1,842	**5,175**	206	19	63	3
Opposing Position	9	37	823	299	**1,319**	15	41	3
Transitions	4	5	289	20	15	**264**	2	0
Conclusion	117	0	619	81	34	5	**4,622**	16
No Tag	3	32	113	4	3	0	13	**81**

Rows denote Annotator 1 and columns denote Annotator 2. Bold values indicate agreement between Annotator 1 and Annotator 2.

The contingency table (Table 6) further explores disagreements. The *Elaboration* tag was the most assigned tag. When there were disagreements, the other annotator most frequently labeled the same sentence as either *Evidence*, *Introduction*, or *Opposing Position*. When the annotators disagreed on an *Evidence* label, the most common label assigned was *Elaboration*. These *Evidence* disagreements could be, in part, due to the familiarity that many of the annotators had with the prompts: They were keen to detect when a student's writing was primarily a summarization of the source passages that accompanied the prompts. Even though the researchers encouraged the annotators to refrain from drawing upon their knowledge of the other sources, it may have been difficult to know where to draw the line between text that was evidence and elaboration, given deep background knowledge of the source prompts. As such, this pattern of disagreements may not generalize to a different set of prompts and annotators.

The label for *Opposing Position* was infrequently assigned, which may be explained in part by the fact that grade 6 students are not instructed to include this element in their essays. When annotators disagreed, one rater most likely assigned a label of *Elaboration*. Two possible reasons for this disagreement between *Opposing Position* and *Elaboration* are the following. First, at times, students can be quite subtle with their opposing ideas, oftentimes blending them together in a same sentence with *Elaboration* in support of a claim. Second, according to the guidelines, any rebuttal to the opposing position should be marked as *Elaboration*, which may have been a rule that was sometimes overlooked by annotators.

Finally, *Transition* tags appeared to be both infrequent and difficult to agree on. *Transitions* are likely infrequent for two key reasons. First, we defined transitions, in part, as any sentence that re-states any part of the introduction. If a sentence that would otherwise be a transition includes *any* new information, it should be marked as *Elaboration*. Second, the *Transition* tag is the last in the hierarchy of tags. That is, if a sentence could be identified as any other writing element, it should be tagged as such.

This analysis identifies potential improvements of the annotator training process to clarify and emphasize aspects of the guidelines. Yet, the agreement indices are sufficiently high to proceed with exploring the extent to which we can use these annotations to differentiate patterns in student writing.

3.2 Annotations analysis

3.2.1 Eight latent classes

We examined six model-fit statistics across the four-class to the ten-class solutions to determine the number of latent class C in the final model (see Appendix C for details on the model fit results). The results indicated that Akaike information criterion (AIC; Akaike, 1974) values consistently decreased as the number of classes increased, while the reductions were minimal after the eight-class model. The values such as Bayesian information criterion (BIC; Schwarz, 1978), sample-size adjusted BIC (ABIC; Sclove, 1987) and consistent AIC (CAIC; Bozdogan, 1987) values were lowest for the eight-class model, indicating the best fit. Approximate weight of evidence (AWE; Banfield and Raftery, 1993) favored the six-class solution. While no one class was unanimously favored across all fit statistics, the eight-class model was chosen as a balanced solution between parsimony and interpretability. Also, the eight-class model repeatedly converged on the same highest log-likelihood in multiple runs.

The eight identified latent classes all exhibited distinct probability profiles across the four features, suggesting eight distinct groups of students' essays. The description for each latent class was derived based on category responses with the highest class-conditional probabilities (the class probabilities for the entire classes can be found in Appendix D). The estimated latent classes are nominal categories with no inherent order. However, for the ease of interpretation, we present the latent classes in the order of their average Purpose Focus and Organization (PFO) and Evidence and Elaboration (EE) scores, as shown in Table 7. We describe each class in more detail below.

Note the patterns that emerge as the classes progress in relation to average rubric scores. First, no class's average essay score falls within the highest rubric category, reflecting how uncommon top-level writing is according to this rubric. However, as average scores increase, there is clear alignment with the qualitative descriptors of each rubric level. For the initial classes, with average scores near level 1, essays typically show "little or no discernible organizational structure" and lack substantive evidence or elaboration, which may be "minimal or absent." As average scores approach level 2, organizational structures begin to emerge, though they often remain inconsistent -- for instance, a controlling idea may appear in a body paragraph rather than the introduction. Evidence and elaboration also begin to emerge. For classes with scores approaching level 3, evidence and elaboration may appear together, suggesting

TABLE 7 Eight estimated latent classes and their average rubric scores and essay length.

Latent class	Description	Avg OS	Avg EES	Avg EL
1 (Body-only with no controlling idea)	Essays composed solely of body paragraphs without any *Controlling Idea* sentence	1.29	1.24	14.9
2 (Controlling-idea only in the introduction)	Essays containing a paragraph composed of a single *Controlling Idea* sentence	1.61	1.43	15.0
3 (Missing introduction or conclusion, but containing controlling idea)	Essays missing an introduction or conclusion paragraph, or missing both, but containing a *Controlling Idea*	1.65	1.47	17.6
4 (Controlling idea in the introduction paragraph with introductory remarks)	Essays containing both a *Controlling Idea* and *Introduction* sentences in an introduction paragraph	1.78	1.51	19.1
5 (Missing or hidden controlling idea)	Essays featuring a conventional "Introduction – Body – Conclusion" structure but missing a *Controlling Idea* sentence or with it hidden in a body paragraph	1.97	1.77	24.5
6 (Controlling idea presented in the conclusion)	Essays featuring a conventional "Introduction – Body – Conclusion" structure with a *Controlling Idea* sentence in the conclusion paragraph	2.08	1.89	26.9
7 (Multiple introduction)	Essays with a conventional "Introduction – Body – Conclusion" structure but containing multiple introduction paragraphs	2.22	2.02	27.0
8 (Conventional structure)	Essays featuring a conventional "Introduction – Body – Conclusion" structure with elaborated argument with evidence	2.24	2.05	27.1

Avg is the average; OS: average organization score; Avg ES: average elaboration and evidence score; Avg EL: average essay length.

integration and indicating a clearer alignment with the criteria at that level. As scores continue to rise through the second, and approaching the third rubric level, the essays become increasingly structured and supported, even showing signs of considering the opposing position.

- Class 1 (Body-only with no Controlling Idea): This class represents 4.7% of essays. The average scores for Purpose, Focus, and Organization (PFO) and Evidence and Elaboration (EE) are 1.29 and 1.24, respectively and are near level 1 of the rubric. According to the rubric, essays at this level may lack a controlling idea and show little to no discernable organizational structure. Consistently, essays in this class have a 78% chance of missing an introduction paragraph, a 100% chance of lacking a *Controlling Idea* sentence, and an 87% chance of omitting a conclusion paragraph.
- Class 2 (Controlling Idea-Only in the Introduction): Accounting for 9.4% of the population, these essays have average scores of 1.61 (PFO) and 1.43 (EE), still near level 1. While most lack strong organization, some include an emerging controlling idea. Notably, these essays have a 91% probability of having an entire introduction paragraph that simply consists of a single controlling idea sentence. This notable characteristic is also reflected in the introduction feature as either having a 70% probability of having an introduction paragraph solely consisting of a controlling idea sentence or having a 30% probability of multiple introduction paragraphs as an artifact of having a paragraph with only a controlling idea sentence. This class also exhibited more than 50% chance of missing a conclusion paragraph, contributing to short length and low scores.
- Class 3 (Missing Introduction and/or Conclusion, but Contains a Controlling Idea): This class includes 22.2% of essays, with slightly higher average scores than Class 2 – indicating a partial structural emergence. These essays have a 79% chance of missing an introduction paragraph, a 100% chance of containing a controlling idea sentence in the body paragraph, and a 65% chance of lacking a conclusion paragraph. While these essays do contain a main claim, they still lack a standard introductory structure where the main claim is naturally integrated with introductory remarks. These limitations may explain their relatively low scores.
- Class 4 (Controlling Idea in the Introduction Paragraph with Introductory Remarks): The fourth class was estimated to account for 22.5% of the essays in the population. These essays average 1.78 for PFO and 1.51 for EE – approaching level 2. They exhibit a strong introductory structure with an 88% probability of containing an introduction paragraph that consists of both introduction and controlling idea sentences. While these essays show a standard introductory structure, they tend to miss a conclusion paragraph with a 43% probability. While the organization structure is developing, the use of evidence and elaboration is not yet emerging.
- Class 5 (Missing or Hidden Controlling Idea): The fifth class of essays, which accounted for 6.5% of the population, showed a 90% chance of having an introduction paragraph, consisting of introduction sentences only, while their controlling idea is either located in the body paragraph (53%) or is missing (47%). The average PFO score of 1.97 suggests emerging structure. Also, unlike the previous classes, this class showed relatively high chances of containing *Elaboration* and *Evidence* sentences together in the body paragraph (52%) and a long conclusion (44%). With an EE score of 1.70, this class shows "uneven, cursory support" that begins to align with rubric expectations.

- Class 6 (Controlling Idea Presented in the Conclusion): Representing 4.3% of essays, this class has average scores that place it at or near level 2. These essays showed an 86% probability that their introduction paragraph consists of only introduction sentences and a 96% that their controlling idea is located the conclusion paragraph. While the structure may present itself as unconventional, the essays received higher essay scores perhaps due to the rhetorical strategy of building up to the controlling idea in the final paragraph.
- Class 7 (Multiple Introductions): This class comprises 3.5% of all essays. All essays included multiple introduction paragraphs, and many feature both evidence and elaboration (50%) and long conclusions (43%). Average rubric scores – 2.22 (PFO) and 2.02 (EE) – are firmly in level 2. These essays show developing structure and support.
- Class 8 (Conventional Structure) The largest class, at 26.8%, features essays with strong alignment to the rubric descriptors. With average scores of 2.24 (PFO) and 2.04 (EE), these essays typically include an introduction with both introductory remarks and a controlling idea (93%), a long conclusion (68%), and integrated evidence and elaboration in the body (49%). Some even feature an opposing position and rebuttal (29%). These well-structured essays receive the highest rubric score across all classes.

Figure 3 displays the posterior probability of latent classes based on Organization rubric scores. The figure suggests significant differences in class probability distributions across the essay scores. For example, among essays receiving the Organization score of 1, the third latent class had the highest probability of occurrence (0.341), followed by the fourth (0.231), second (0.161), and first (0.129). The probability of these essays belonging to the eighth class was only 0.059. However, these trends noticeably changed with higher Organization scores. As the score increased to 4, the proportions of the third and fourth classes, characterized by a lack of structure and organization, drastically decreased, respectively, to 0.055 (third) and 0.041 (fourth). The proportion of the eighth class, which includes essay with a standard format and sophisticated elaboration, on the other hand, grew with higher scores. In particular, the probabilities for the essays scoring 3 and 4 falling into the eighth class were substantially higher than those for the other classes, marking 0.553 (score 3) and 0.664 (score 4), respectively. This result suggests meaningful differences in rubric scores across different writing patterns reflected in latent classes.

4 Discussion and limitations

Our findings demonstrate that carefully crafted annotation guidelines, aligned with established rubrics and standards can provide a reliable framework for detecting finer-grained argumentation elements at scale. Just at rubrics standardize holistic scoring, these guidelines enable consistent identification of sentence-level argumentative elements across large numbers of essays, when combined with high-quality training materials and methods. The documented development process, annotator training procedures, and agreement statistics support the use of this approach. Importantly, our analysis reveals that these fine-grained annotations not only identify distinct writing patterns but also align meaningfully with rubric score points, suggesting they effectively complement traditional scoring methods. As such, the annotations and LCA results offer ways to

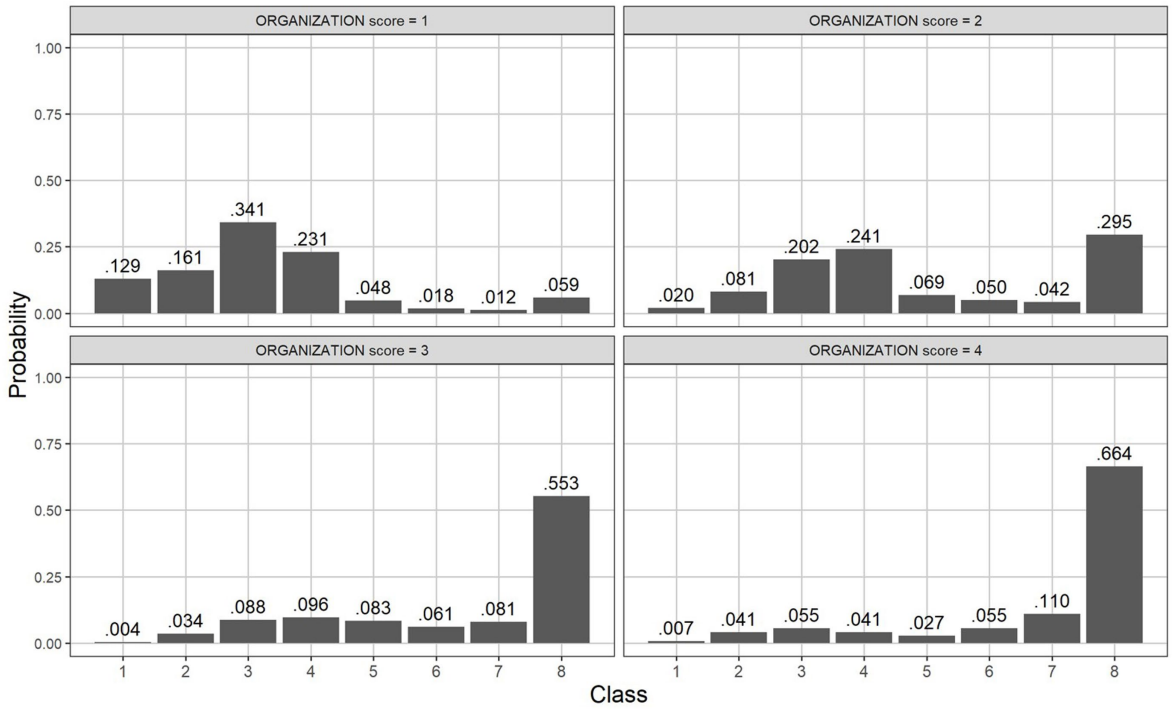

FIGURE 3
Score-conditional posterior probabilities across the eight identified classes.

better understand organizational patterns in writing and how those patterns are aligned with the rubric scores.

While this study advances our understanding of scalable writing assessment, several limitations warrant consideration. First, our focus on 6th-8th grade argumentative essays, while substantive, represents only one segment of academic writing. Future work should extend these guidelines to different grade levels and genres, including explanatory and narrative writing. Even within the 6–8 grade band, aspects of the annotation guidelines merit further discussion, including the use of a hierarchical system for assigning a single annotation label, the decision to annotate at the sentence level, and the providing clarity of the *Transitions* label.

A key design decision in this study was the reliance on a hierarchy to assign a single tag to each sentence. This was intended to reduce the cognitive load on annotators while ensuring the most critical argumentative components, such as the controlling idea or opposing position, were prioritized. However, this approach has limitations, particularly when applied to compound sentences that express multiple functions. The restriction to one tag per sentence may obscure the presence of meaningful argumentative elements, contributing to information loss and potentially reducing inter-rater agreement.

This challenge is closely tied to the choice of sentence-level annotation rather than clause-level. While annotating at the sentence level offers benefits–including greater efficiency, reduced annotator burden, and more straightforward and interpretable agreement calculations -- at times, it may oversimplify student writing. Notably, Stab and Gurevych (2014) found that 5.6% of sentences contained multiple argumentative labels in their corpus, indicating that while the issue is relatively infrequent, it remains important to consider. One way to mitigate these concerns, without shifting to clause-level annotation, is to replace the hierarchical structure with a more flexible tagging approach. This could involve introducing labels for common compound sentence structures, such as those that combine evidence and elaboration. By accommodating multi-functional sentences, such an approach could retain the practical benefits of sentence-level annotation while better capturing the complexities of student writing.

Of the seven annotation labels included in guidelines, *Transitions* emerged as the most challenging to define and apply consistently. According to the rubrics guiding this work, transitions are strategies used to "clarify the relationships among ideas." However, this concept is difficult to operationalize. To avoid overemphasizing the presence of transitional words alone, we adopted a broader definition that includes such strategies as reiterating sub-claims from the introduction, signaling what is coming next at the end of a paragraph, or using sentences that introduce no new information but serve to enhance coherence. While this broader view aims to capture transitions at the essay level, it may unintentionally blur distinctions at the paragraph level, where transitional and topic sentence may be the same. Furthermore, the subtlety and complexity of transitional strategies, especially when effectively executed, suggest the need for a more targeted approach. The next step would be to develop a dedicated annotation scheme focused specifically on defining and identifying transitional strategies throughout an essay. In the short term, finding ways to adopt and clearly communicate a more precise term for these strategies, as defined in our annotation guidelines, may help reduce confusion. To end, a final key limitation inherent in our approach is that the seven argumentation elements, though crucial components identified in rubrics, represent necessary but not sufficient conditions for effective argumentation. The presence of these elements alone cannot guarantee a compelling argument–factors such as stylistic choices, evidence quality, idea cohesion, and controlling idea strength all contribute to overall writing quality. However, even acknowledging that such annotations do not capture all aspects of writing, our findings offer promise that these annotations, when systematically applied and interpreted, can still provide valuable guidance to students and educators during the writing and revising process. Future work should investigate how this detailed feedback can be most effectively integrated into classroom practice and automated assessment systems.

Data availability statement

The datasets presented in this article are not readily available due to data sharing agreements that have strict stipulations on data disclosure.

Author contributions

AmB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. SH: Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing, Conceptualization. SW: Conceptualization, Investigation, Methodology, Supervision, Writing – original draft. AlB: Investigation, Methodology, Supervision, Writing – review & editing, Conceptualization. FR: Investigation, Methodology, Supervision, Writing – review & editing, Conceptualization. SL: Investigation, Methodology, Supervision, Writing – review & editing, Conceptualization, Resources.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We would like to thank Mackenzie Young and Ben Godek for their contributions to data collection and analysis.

Conflict of interest

AmB, SH, SW, AlB, FR, and SL were employed by Cambium Assessment, Inc.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1569529/full#supplementary-material

References

Akaike, H. (1974). A new look at the statistical model identification problem. *IEEE Trans. Autom. Control*, 19, 716.

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College teaching*, 53, 27–31.

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). Discrete multivariate analysis: Theory and practice: Springer Science & Business Media.

Brookhart, S. M., and Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educ. Rev.* 67, 343–368. doi: 10.1080/00131911.2014.929565

Banfield, J. D., and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803–821.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345–370.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Council of Chief State School Officers & National Governors Association Center for Best Practices. (2010). Common Core state standards for English Language Arts & Literacy in history/social studies, science, and technical subjects. Available online at: https://learning.ccsso.org/wp-content/uploads/2022/11/ELA_Standards1.pdf (Accessed May 22, 2025).

Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Benner, M., and Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assess. Writ.* 54:100667. doi: 10.1016/j.asw.2022.100667

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382. doi: 10.1037/h0031619

Florida Department of Education. (2025). Writing assessments. Florida Department of Education. Available online at: https://www.fldoe.org/accountability/assessments/k-12-student-assessment/writing.stml (Accessed January 31, 2025)

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., and Gurevych, I. (2018): The INCEpTION platform: machine-assisted and knowledge-oriented interactive annotation. In Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, New Mexico, USA.

Krippendorff, K. (2004). Measuring the reliability of qualitative text analysis data. *Qual. Quant.* 38, 787–800. doi: 10.1007/s11135-004-8107-7

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. In Biometrics 33, 159–174. doi: 10.2307/2529310

Lazarsfeld, P. F. (1968). Latent structure analysis.

Linzer, D. A., and Lewis, J. B. (2011). poLCA: an R package for polytomous variable latent class analysis. *J. Stat. Softw.* 42, 1–29. doi: 10.18637/jss.v042.i10

Nguyen, H., and Litman, D. (2018). Argument mining for improving the automated scoring of persuasive essays. In Proceedings of the AAAI Conference on Artificial Intelligence

Ormerod, C., Burkhardt, A., Young, M., and Lottridge, S. (2023). Argumentation element annotation modeling using XLNet. arXiv preprint arXiv, 2311.06239.

Smarter Balanced Assessment Consortium. (2022). Elaboration guidelines grades 6–11. Available online at: https://portal.smarterbalanced.org/wp-content/uploads/Elaboration-Guidelines-Gr6-11.pdf (Accessed May 22, 2025).

Smarter Balanced Assessment Consortium. (n.d.). Argumentative performance task writing rubric (grades 6–11). Smarter balanced. Available online at: https://portal.smarterbalanced.org/library/en/performance-task-writing-rubric-argumentative.pdf (Accessed January 31, 2025)

Stab, C., and Gurevych, I. (2014). Annotating argument components and relations in persuasive essays. In Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers (pp. 1501–1510)

Stab, C., and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Comput. Linguist.* 43, 619–659. doi: 10.1162/COLI_a_00295

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52, 333–343.

Texas Education Agency. (2022). STAAR argumentative/opinion writing rubric: grades 6–English II. Texas Education Agency. Available online at: https://tea.texas.gov/student-assessment/staar/staar-6-english-ii-arg-opinion-rubric.pdf (Accessed January 31, 2025)

Toulmin, S. E. (2003). The uses of argument. Cambridge, England: Cambridge University Press.

Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educ. Meas. Issues Pract.* 31, 2–13. doi: 10.1111/j.1745-3992.2011.00223.x