

Does “Effective” Mean the Same Across Discourse Types? Decomposing Discourse Effectiveness into Elaboration and Length-Robust Structure

Yi Gui

Measurement Incorporated
ygui@measinc.com

Abstract

Automated writing evaluation often treats discourse-effectiveness labels as prediction targets, but predictive accuracy cannot establish that such labels support stable construct interpretation across discourse functions. Using 144,289 discourse segments from 15,593 PERSUADE 2.0 essays, this study examined whether discourse effectiveness has an invariant linguistic realization across discourse types, or is conditioned by rhetorical function—and how much of that conditioning reflects elaboration rather than length-robust structure. Segments were represented with 191 traditional NLP features, reduced to eight principal components, and analyzed with discourse type $\times$ effectiveness contingency tests, MANOVA, component-level ANOVAs, random-intercept mixed-effects models, and length decomposition. Discourse type and effectiveness were associated, and their interaction appeared across multiple components and persisted after accounting for essay-level clustering. The interaction was largely elaboration-borne: the dominant component was strongly length-linked ($R^2 \approx .62$ with word count), and after length adjustment its interaction effect fell by over 80%, while the multivariate interaction was attenuated by 39–47% yet remained and cross-type reversal patterns persisted. These findings do not invalidate the PERSUADE 2.0 labels; rather, discourse effectiveness is type-conditioned and largely elaboration-borne in this representation—a fragile basis for context-free diagnosis—though a smaller length-robust residual persists.

1 Introduction

AI-enabled writing assessment is typically evaluated through score agreement, predictive accuracy, and operational scoring criteria, and those criteria remain important for automated essay scoring (AES) and automated writing evaluation (AWE) (Deane, 2013; Ke & Ng, 2019; Ramineni &

Williamson, 2013). However, the measurement value of an AI-supported representation also depends on the construct it is assumed to encode and on the validity of the interpretations drawn from observed performances (Kane, 2013; Messick, 1995). Annotation labels, prediction targets, measurement constructs, and feedback categories are therefore not interchangeable uses of the same information: a label may be easy to predict statistically and still be difficult to interpret educationally if its linguistic meaning changes across contexts. For discourse-level analysis this is especially salient, because the same evaluative label is applied to segments serving different rhetorical functions within an essay.

The PERSUADE 2.0 corpus (Crossley et al., 2024) provides discourse segment boundaries, discourse-type labels, and discourse-effectiveness labels at a scale relevant to computational writing assessment, making it possible to ask not only whether a model can recover a label, but what that label means linguistically. At first glance, the three effectiveness categories—Ineffective, Adequate, and Effective—invite an ordinal interpretation under which more effective segments differ from less effective segments in relatively stable ways across discourse functions. That assumption is not guaranteed. Argumentation research treats discourse components as serving distinct functional roles (Stab & Gurevych, 2017), so the linguistic profile of an effective Evidence segment may not match that of an effective Lead, Claim, Rebuttal, or Concluding Statement.

A label’s observable meaning can also be carried by features only indirectly related to what it is intended to capture. One such feature is elaboration: more developed segments contain more words, more sentence-to-sentence overlap, and—mechanically—lower type-token ratios. Elaboration is plausibly construct-relevant, because well-developed evidence or a fully realized conclusion is often judged more effective, yet it is entangled with

measurement artifacts, since raw lexical-diversity indices decline with length. To the extent that the signal for an effectiveness label is dominated by elaboration, a context-free reading of that label becomes harder to defend: the representation may be responding less to function-specific quality than to how much a writer produced.

The present study therefore approached discourse effectiveness as a construct-interpretation question with two parts. First, within an interpretable traditional NLP feature space, does effectiveness support a single invariant linguistic realization across discourse types, or is its realization conditioned by discourse type? Second, if it is conditioned, how much of that conditioning is attributable to segment elaboration, and how much survives as length-robust linguistic structure? Treating elaboration as double-edged allows the analysis to separate type-conditioned structure that reflects how much a writer elaborated from type-conditioned structure that does not.

This framing yields a more precise contribution than a demonstration that effectiveness merely “differs” across discourse types. By decomposing the discourse type $\times$ effectiveness interaction into an elaboration-linked component and a length-robust residual, the study clarifies what an effectiveness label indexes at the segment level, and what this implies for systems that assign, predict, or give feedback on effectiveness as if its meaning were stable across discourse functions (Kane, 2013; Messick, 1995).

2 Related Work

Research on AES and AWE has historically emphasized score prediction, rater agreement, and generalization across prompts or domains (Deane, 2013; Ke & Ng, 2019; Ramineni & Williamson, 2013), while raising a recurring construct question: the features most accessible to computational modeling are not always those that best represent the intended writing construct (Deane, 2013; Messick, 1995). This matters most when labels are used for interpretation, diagnosis, and feedback rather than prediction alone. Studies of formative systems such as eRevise, with argument-based validation work on automated feedback, emphasize that feedback must rest on a defensible validity argument and reflect meaningful aspects of the writing construct (Correnti et al., 2022; Ranalli et al., 2017; Wang et al., 2020), while reviews of AWE show rapid ex-

pansion alongside uneven validation and persistent construct-representation challenges (Nunes et al., 2022; Shi & Aryadoust, 2024; Stevenson & Phakiti, 2014).

This aligns with measurement-oriented work arguing that assessment should be built around explicit links between observed evidence, score meaning, and the claims made from performances (Kane, 2013; Messick, 1995; Mislevy et al., 2003), which implies that labels should not be treated as invariant traits: the relation between text features, intermediate representations, and final judgments must be examined empirically. Trait- and rubric-based approaches insert diagnostic structure between essay input and final evaluation, but still often assume that a target label carries a stable meaning across contexts.

The present study extends this concern to discourse-level effectiveness. Argumentation research has shown that persuasive essays contain functionally distinct discourse components (Stab & Gurevych, 2017), yet little work has asked whether effectiveness itself has a stable linguistic meaning across those functions. If it is expressed differently in different discourse types, a global effectiveness model may conflate distinct quality criteria while still appearing successful by prediction metrics.

3 Methods

3.1 Data and Sample

The study used the PERSUADE 2.0 corpus (Crossley et al., 2024). The analytic unit was the discourse segment rather than the essay. Each observation corresponded to a labeled segment with two focal annotation variables: discourse type and effectiveness. Discourse-type labels identified the rhetorical function of the segment (Lead, Position, Claim, Counterclaim, Rebuttal, Evidence, Concluding Statement), consistent with the broader use of discourse and argument components in computational analyses of persuasive essays (Stab & Gurevych, 2017). Effectiveness labels represented three ordered rating categories assigned at the segment level (Ineffective, Adequate, Effective).

The data file was filtered to retain only segments with non-empty text and non-missing effectiveness labels, yielding 144,289 segments nested within 15,593 essay-level groups defined by the stable essay identifier in the corpus metadata. Because the study addressed a construct question rather than a predictive one, the design emphasized interpretable

representations, effect sizes, and interaction structure rather than accuracy optimization.

3.2 Feature Representation

The feature space was designed to make the linguistic evidence for effectiveness visible and interpretable at the segment level. This study does not directly use TAACO, Coh-Metrix, or TAALES software outputs; features were operationalized through custom Python implementations following documented operational definitions, and those labels serve only as conceptual reference points. Six families were represented: lexical diversity (lemma type–token ratio, MATTR); cohesion and overlap (adjacent and skip-one sentence or paragraph lemma overlap); semantic similarity (LSA, LDA, Word2Vec similarity across adjacent units); readability and surface structure (Flesch–Kincaid, Flesch Reading Ease, Gunning Fog, punctuation and length counts); syntax (sentence length, dependency depth, clause count, subordination ratio, POS ratios from spaCy parses); and discourse connectives (causal, additive, logical, temporal, contrastive, and coordinating rates from predefined lexicons). The full feature dictionary, package sources, and preprocessing notes appear in Supplementary Appendix A and Table S1.

Extraction produced 191 numeric features per segment; its role was evidentiary rather than exhaustive. Features with zero or near-zero variance were removed (13 features; 178 retained), and the retained features were z -standardized so that counts, proportions, ratios, readability indices, and similarity measures contributed on a comparable scale. Missingness in the final analytic matrix was negligible.

3.3 Latent Dimension Extraction

Modeling all 178 features directly would have created redundancy among closely related measures, a large multiple-testing burden, and interpretive difficulty. Principal components analysis (PCA) was therefore used to summarize the space into continuous latent dimensions that preserved dominant covariance structure while remaining interpretable (Jolliffe & Cadima, 2016). Highly correlated features were first pruned at $|r| > .90$, removing 66 features and leaving 112 for extraction. Retention combined parallel analysis on a stratified sample of 12,000 segments (up to 4,000 per effectiveness label) with 40 random reference iterations (Horn, 1965), a scree-elbow criterion (Cattell, 1966), and

interpretability. Parallel analysis indicated substantial high-dimensional structure, whereas the scree plot suggested a more compact set of dominant dimensions; the first eight components were retained, accounting for 50.6% of total variance. Supplementary Appendix B reports explained variance, loading anchors, the full loading matrix, and scree/loading figures.

Components were interpreted from dominant loading families. The analyses below focus on PC1 (Cohesion/overlap vs. Lexical diversity), PC2 (Cohesion/overlap, an internal configuration of cohesion features rather than a one-directional coherence scale), PC3 (Syntactic complexity vs. POS distribution), and PC6 (Connectives/discourse relations vs. Cohesion/overlap). These labels are interpretive shorthand rather than psychometric constructs; full loading evidence appears in Tables S2–S4.

3.4 Analytic Models

A contingency table of discourse type by effectiveness was evaluated on the full dataset with Pearson’s chi-square test; because of the large sample and asymmetric label distribution, effect magnitude was summarized with Cramér’s V and standardized residuals were computed for each cell.

A multivariate analysis of variance (MANOVA) then tested whether the joint latent profile differed by discourse type, effectiveness, and their interaction, with the eight component scores as dependent variables. Because the 7×3 design was highly unbalanced, MANOVA was estimated on a balanced-cell sample drawing up to 2,000 segments per cell ($N = 30,718$), with Pillai’s trace as the primary summary. To locate the interaction, two-way ANOVAs of the form $PC_k \sim \text{discourse type} + \text{effectiveness} + \text{discourse type} \times \text{effectiveness}$ were fit for each component on the same sample, with partial η^2 from Type III sums of squares; at this sample size significance alone is uninformative, so interpretation emphasized the relative magnitude of interaction effects.

Because segments were nested within essays, random-intercept models of the form $PC_k \sim \text{discourse type} \times \text{effectiveness} + (1 | \text{essay group})$ were estimated by maximum likelihood on the full 144,289-segment dataset for the four components with the largest interaction effects (PC1, PC2, PC3, PC6), yielding term-level tests, predicted means, and intraclass correlation coefficients (ICCs) as a robustness check on the balanced-sample results.

3.5 Length Decomposition

Because several traditional NLP features are sensitive to segment length—word count directly, and raw type–token ratios mechanically—the dominant component was expected to carry an elaboration signal. The interaction was therefore re-estimated under two adjustments. In component residualization (B1a), each component score was regressed on a natural-spline function of log segment length and the analyses repeated on the residualized scores. In length-purged re-extraction (B1b), word and sentence counts were removed and every remaining feature was residualized on the same length function before re-standardization, correlation pruning, and re-extraction, after which the full pipeline was repeated. Two sensitivity analyses were added: one removing raw type–token-ratio indices, and one residualizing features on length within each discourse type (Table S12).

4 Results

4.1 Descriptive Structure

Segments were unevenly distributed across discourse types, with Claim (34.80%) and Evidence (31.67%) dominating and Rebuttal (3.01%) least common, and the effectiveness distribution was highly imbalanced (4.48% Ineffective, 78.48% Adequate, 17.04% Effective; Table 1). Despite this skew, effectiveness composition varied by type: Lead had the highest proportion of Effective labels (21.39%), followed by Concluding Statement (18.63%) and Evidence (18.56%), whereas Position had the lowest (12.07%) and one of the strongest concentrations in Adequate (84.88%). Evidence had the highest Ineffective proportion (6.91%).

The contingency analysis confirmed this pattern: $\chi^2(12) = 1642.97$, $p < .001$, with a modest association magnitude (Cramér's $V = 0.075$). At this sample size the standardized residuals are more informative than the p -value (Figure 1). The strongest positive residual was Evidence $\times$ Ineffective (24.52), with Evidence $\times$ Effective also overrepresented (7.89); Position was overrepresented in Adequate (8.97) but underrepresented in Effective (−14.94), and Claim showed a similar pull toward the middle category (Ineffective −16.31). Lead $\times$ Effective (10.17) and Concluding Statement $\times$ Effective (4.49) were overrepresented, Counterclaim $\times$ Effective underrepresented (−4.77), and Rebuttal close to independence.

4.2 Latent Dimensions and Multivariate Structure

The retained components functioned as summary coordinates compressing recurring structure in a feature set spanning lexical diversity, cohesion and overlap, syntactic complexity, and discourse markers; none was equivalent to a single global effectiveness scale. All three MANOVA terms were statistically significant, but relative magnitudes are more informative than p -values (Table 2): Pillai's trace was 0.444 for discourse type, 0.142 for effectiveness, and 0.093 for their interaction (all $p < .001$). The largest multivariate effect was discourse type, indicating that rhetorical function accounted for a substantial share of structure in the latent space. Most importantly, the interaction was robust: the multivariate profile associated with effectiveness differed across discourse types.

4.3 Component-Level Patterns

Ranking the component-level interaction terms by effect size located the multivariate interaction (Table 3): the strongest emerged for PC1 (partial $\eta^2 = 0.039$), followed by PC3 (0.014), PC2 (0.011), and PC6 (0.008). PC1 also showed by far the largest discourse-type main effect ($\eta_p^2 = 0.335$) and a smaller effectiveness main effect ($\eta_p^2 = 0.068$), suggesting a broad axis of discourse variation on which effectiveness distinctions were important but not parallel across types.

Cell means showed that Effective segments generally had higher PC1 scores than Ineffective segments, but the magnitude and internal ordering differed substantially by discourse type. Lead moved from −0.80 (Ineffective) to 7.19 (Adequate) to 21.05 (Effective), and Concluding Statement from −3.32 to 11.81 to 25.67, while Evidence separated sharply between Effective (28.08) and the other categories (13.56; 14.43). Position and Claim were weaker and non-monotonic: for Position, Adequate (−14.90) fell below Ineffective (−11.99), with Effective highest (−7.62); Claim showed the same shape (−15.21, −13.56, −12.47).

PC2, PC3, and PC6 departed further from global invariance. On PC2, higher scores were often associated with *lower* effectiveness, especially for Evidence and Concluding Statement. On PC3, the effectiveness contrast reversed across types: Effective was highest for Position and Counterclaim but not for Rebuttal, Evidence, or Concluding Statement. PC6 likewise showed crossed patterns, with

Effective exceeding Ineffective for Lead, Evidence, and Concluding Statement but not for the other types. Figure 2 summarizes these monotonic, non-monotonic, and reversed profiles; further plots appear in Figures S3–S5.

4.4 Mixed-Effects Models

ICCs indicated non-trivial essay-level clustering: 0.103 (PC1), 0.210 (PC2), 0.227 (PC3), and 0.058 (PC6), so roughly 6%–23% of variance was between essays (Tables S8–S11).

The mixed-effects results were highly consistent with the earlier analyses (Table 4): for all four dimensions, discourse type, effectiveness, and their interaction were significant (interaction $F = 235.43, 60.37, 84.68, \text{ and } 52.66$ for PC1, PC2, PC3, and PC6; all $p < .001$). Predicted means again showed that the size and direction of the effectiveness contrast varied by discourse function: on PC1, Effective exceeded Ineffective in every type but by very different margins; on PC2, Effective was generally lower than Ineffective, especially for Evidence and Concluding Statement; and on PC3 and PC6 the direction continued to reverse. Modeling essay-level dependency did not remove the interaction.

4.5 Elaboration Versus Length-Robust Structure

Word count alone accounted for most variance in the leading component ($R^2 \approx .62$), so the decomposition examined how much of the interaction reflected elaboration rather than length-robust structure (Table 5).

Both adjustments attenuated the interaction without eliminating it. At the multivariate level it was reduced by 39%–47% (Pillai's trace from $\approx .093$ to $\approx .049$ – $.057$) but remained reliable. The reduction was concentrated in the dominant, length-linked component, whose interaction effect fell by about 83%–86% ($\eta_p^2 \approx .039$ to $\approx .005$ – $.007$), after which it no longer carried a meaningful share of the type-conditioned structure. By contrast, a paragraph-level cohesion dimension was largely length-robust, and non-parallel and reversed cell-mean profiles persisted across components rather than disappearing. Sensitivity checks indicated that the attenuation reflected segment size broadly rather than a mechanical type–token-ratio artifact: removing raw type–token-ratio indices left the interaction essentially unchanged, whereas adjusting for length within each discourse type produced the

largest attenuation yet still did not remove it.

Much of the apparent type-conditioning is therefore carried by how much a segment elaborates, and that coupling itself varies across discourse types, while a smaller length-robust residual remains.

5 Discussion

5.1 Construct Interpretation

The primary contribution is construct evidence for discourse-level AWE. Within the feature space analyzed here, discourse effectiveness did not behave as a globally invariant label: its observable linguistic realization was conditioned by discourse type. A substantial part of that conditioning is carried by segment elaboration, the strength of the elaboration–effectiveness relationship itself differs across discourse functions, and after elaboration is removed a smaller but genuine length-robust residual remains. The same label therefore corresponds to different linguistic tendencies depending on rhetorical function, and much of that difference is a difference in elaboration.

This does not rest on statistical significance, which at this sample size is uninformative, but on the pattern and the decomposition: the dominant linguistic axis is largely an elaboration dimension; the elaboration–effectiveness coupling varies by type; and the interaction is attenuated but not eliminated by length adjustment, with non-parallel and reversed profiles persisting. Treating elaboration as double-edged is central to this reading. Elaboration is plausibly construct-relevant, because more developed evidence or a more fully realized conclusion is often judged more effective; yet it is entangled with mechanical artifacts, such as the decline of raw type–token ratios with length, that would mislead if read as lexical quality.

The interpretation should remain narrow. The results do not imply that raters erred, that the PER-SUADE 2.0 labels are invalid, or that effectiveness lacks ordinal meaning within discourse types. They indicate that the observable meaning of an effectiveness label is type-conditioned and substantially elaboration-borne—a fragile basis for a single context-free diagnostic reading. The labels may be useful annotation categories, but their use as measurement constructs or feedback targets requires additional validity warrants and explicit attention to elaboration.

5.2 Implications for Modeling and Feedback

A single pooled effectiveness classifier may recover labels well while conflating two things: discourse-specific quality criteria and the amount a segment elaborates. Because the elaboration–effectiveness relationship varies across discourse functions, a model that does not represent discourse type can absorb much of the signal through length-correlated features and still appear successful by prediction metrics, while offering little purchase on why a segment was judged effective. Prediction accuracy alone therefore cannot establish that a label carries the same construct meaning across discourse types. The results motivate representations that are both type-aware and length-aware, rather than letting segment length stand in for quality.

The same holds for feedback. A system should consider what discourse function a segment serves and what counts as effectiveness for that function, and it should be length-aware: a system that responds to length-sensitive features risks telling a writer that a long, developed segment has poor lexical variety when the low type–token ratio is a mechanical consequence of length, or rewarding mere length as quality. For some discourse functions the contrast between lower and higher effectiveness was expressed through length-robust dimensions such as paragraph-level cohesion; for others it reflected elaboration. Useful type-specific feedback would target the former and treat the latter carefully—encouraging development where elaboration is lacking, without equating length with quality.

5.3 Limitations and Future Directions

The conclusions are claims about the analyzed representation rather than about every possible one: traditional NLP features capture meaningful structure but under-represent semantic, pragmatic, and contextual information. Relatedly, all features were computed on each segment in isolation, whereas human labels were assigned with the surrounding essay available to raters. A segment-isolated representation cannot capture how well a segment fits its rhetorical context, so segment length may partly stand in for contextual fit that the present features do not observe. The retained dimensions are also data-driven summaries rather than observed constructs, so their interpretation remains approximate, especially for PC2 and PC6; the analyses were exploratory and warrant replication; the mixed-effects

models used random intercepts rather than slopes; and the study did not implement or test a feedback system.

These limitations point to next steps: replicating the analyses with richer representations, including contextual embeddings; comparing pooled global effectiveness models with type-conditioned models; and, most directly, adding full-essay context measures and re-running the decomposition, which would clarify how much of the effectiveness signal is elaboration, how much is length-robust within-segment structure, and how much is contextual. It may also be more informative to ask whether effectiveness behaves ordinally within each discourse type than across all functions at once.

6 Conclusion

Across descriptive, multivariate, component-level, mixed-effects, and length-decomposition analyses, discourse effectiveness is better understood as discourse-type-conditioned than as a uniform global construct, and much of that conditioning is carried by elaboration rather than length-robust structure. The study does not show that the labels are invalid or that effectiveness lacks ordinal structure; it raises a narrower construct question. Because scoring and feedback systems often treat labels as if their meaning were stable across contexts, a single context-free effectiveness target may both obscure the discourse-specific criteria needed for diagnosis and reward length as if it were quality. More defensible systems will likely need to represent effectiveness together with discourse type, to separate elaboration from length-robust structure, and, ideally, to expose interpretable intermediate linguistic profiles.

References

- [1] Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276.
- [2] Correnti, R., Matsumura, L. C., Wang, E. L., Litman, D., & Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Computers and Education Open*, 3, Article 100084.
- [3] Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227–1237.
- [4] Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Benner, M., & Boser, U. (2024). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing*, 61, Article 100865.

- [5] Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1), 7–24.
- [6] Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.
- [7] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), Article 20150202.
- [8] Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- [9] Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6300–6308).
- [10] Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*, 102(2), 333–349.
- [11] Kyle, K., Crossley, S. A., & Berger, C. M. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046.
- [12] McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
- [13] Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- [14] Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- [15] Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2022). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, 38(2), 599–620.
- [16] Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1), 25–39.
- [17] Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1), 8–25.
- [18] Shi, H., & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, 36(2), 187–209.
- [19] Stab, C., & Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3), 619–659.
- [20] Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing*, 19, 51–65.
- [21] Wang, E. L., Matsumura, L. C., Correnti, R., Litman, D., Zhang, H., Howe, E., Magooda, A., & Quintana, R. (2020). eRevis(ing): Students' revision of text evidence use in an automated writing evaluation system. *Assessing Writing*, 44, Article 100449.

Data and Code Availability

The source corpus is PERSUADE 2.0; this study did not redistribute student text. Supplementary materials include the feature dictionary, PCA loadings, ANOVA and mixed-effects results, and additional figures. Subject to data-use restrictions, analysis scripts and model specifications will be released upon publication.

Tables and Figures

Table 1: Distribution of Discourse Type and Effectiveness

Discourse type	Ineffective	Adequate	Effective	Total
Lead	364 (3.9%)	6,949 (74.7%)	1,990 (21.4%)	9,303
Position	470 (3.0%)	13,087 (84.9%)	1,861 (12.1%)	15,418
Claim	1,475 (2.9%)	40,561 (80.8%)	8,172 (16.3%)	50,208
Counterclaim	205 (3.5%)	4,771 (82.0%)	841 (14.5%)	5,817
Rebuttal	211 (4.9%)	3,406 (78.5%)	720 (16.6%)	4,337
Evidence	3,156 (6.9%)	34,063 (74.5%)	8,483 (18.6%)	45,702
Concluding Statement	581 (4.3%)	10,407 (77.1%)	2,516 (18.6%)	13,504
Total	6,462 (4.5%)	113,244 (78.5%)	24,583 (17.0%)	144,289

Table 2: MANOVA Results (balanced inference sample, $N = 30,718$)

Effect	Pillai's trace	Wilks' Λ	F	Num. df	Den. df	p
Discourse type	0.444	0.581	306.26	48	184,170	< .001
Effectiveness	0.142	0.859	292.94	16	61,382	< .001
Discourse type $\times$ effectiveness	0.093	0.909	30.16	96	245,576	< .001

Table 3: Two-Way ANOVA Results for Selected Principal Components

Component (label)	Effect	F	Partial η^2	p
PC1: Cohesion/overlap vs. Lexical diversity	Discourse type	2579.86	0.335	< .001
	Effectiveness	1124.61	0.068	< .001
	Discourse type $\times$ effectiveness	102.82	0.039	< .001
PC2: Cohesion/overlap	Discourse type	41.84	0.008	< .001
	Effectiveness	131.27	0.008	< .001
	Discourse type $\times$ effectiveness	27.83	0.011	< .001
PC3: Syntactic complexity vs. POS distribution	Discourse type	152.19	0.029	< .001
	Effectiveness	31.53	0.002	< .001
	Discourse type $\times$ effectiveness	35.89	0.014	< .001
PC6: Connectives/discourse relations vs. Cohesion	Discourse type	54.03	0.010	< .001
	Effectiveness	1.70	0.000	.183
	Discourse type $\times$ effectiveness	20.79	0.008	< .001

Table 4: Mixed-Effects Model Summary (full sample; random intercept for essay group)

Component	Label	ICC	Discourse type F	Effectiveness F	Interaction F	Interaction p
PC1	Cohesion/overlap vs. Lexical diversity	0.103	15654.79	158.72	235.43	< .001
PC2	Cohesion/overlap	0.210	535.47	17.16	60.37	< .001
PC3	Syntactic complexity vs. POS distribution	0.227	746.27	236.17	84.68	< .001
PC6	Connectives/discourse relations vs. Cohesion	0.058	364.82	80.16	52.66	< .001

Table 5: Decomposition of the Discourse Type $\times$ Effectiveness Interaction Under Length Adjustment

Metric	Original	B1a (adj.)	B1b (purged)
Interaction (Pillai's trace)	.093	.049	.057
Top-component interaction (η_p^2)	.039	.005	.007
Pillai reduction vs. original	—	47%	39%
Reversing components (of 4)	2	—	3

Note. B1a residualized each component score on segment length; B1b removed length-linked variance from all features before re-extracting components. Pillai values are the discourse type $\times$ effectiveness interaction; η_p^2 is the partial η^2 of the strongest component. Em dashes denote values not applicable.

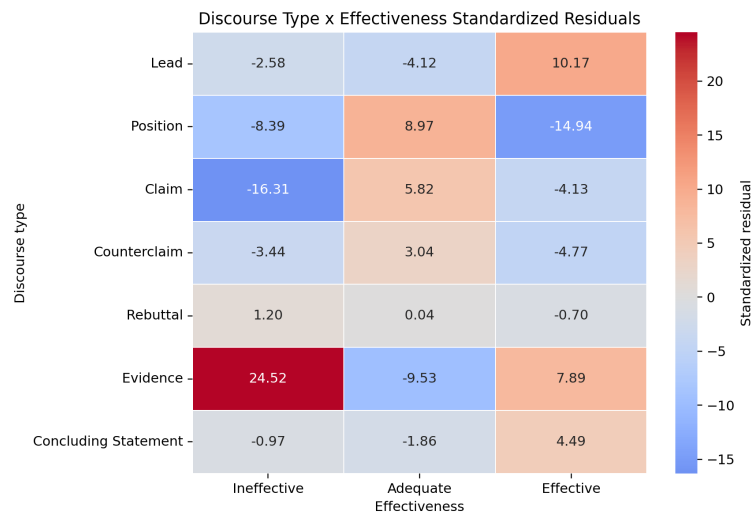


Figure 1: Standardized residuals for discourse type by effectiveness.

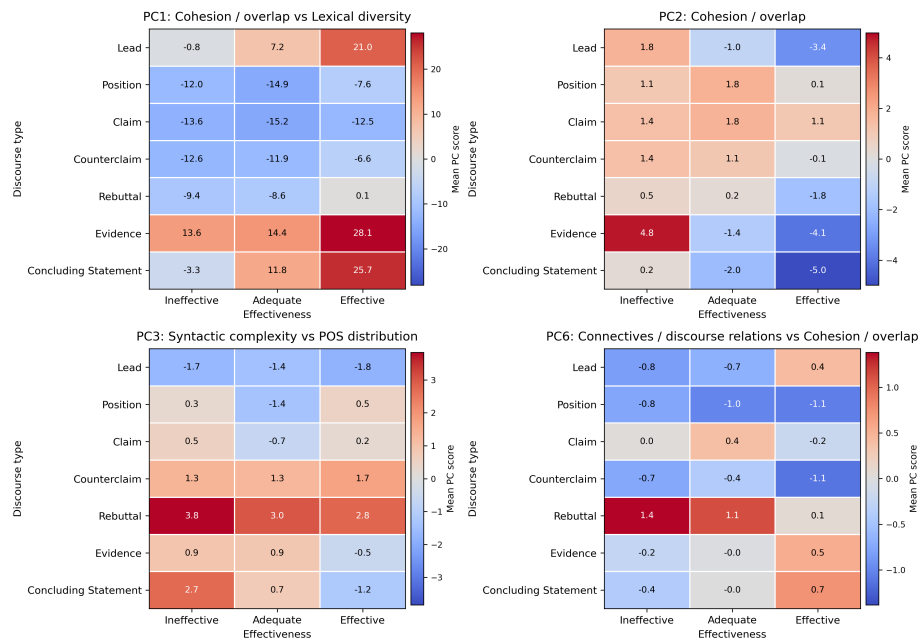


Figure 2: Mean scores for selected principal components by discourse type and effectiveness.