The Impact of an NLP-Based Writing Tool on Student Writing

Karthik Sairam and **Amy Burkhardt** and **Susan Lottridge**
Cambium Assessment
{karthik.sairam, amy.burkhardt, susan.lottridge}@cambiumassessment.com

Abstract

We present preliminary evidence on the impact of a NLP-based writing feedback tool, Write-On with Cambi! on students' argumentative writing. Students were randomly assigned to receive access to the tool or not, and their essay scores were compared across three rubric dimensions; estimated effect sizes (Cohen's d) ranged from 0.25 to 0.26 (with notable variation in the average treatment effect across classrooms). To characterize and compare the groups' writing processes, we implemented an algorithm that classified each revision as Appended (new text added to the end), Surface-level (minor within-text corrections to conventions), or Substantive (larger within-text changes or additions). We interpret within-text edits (Surface-level or Substantive) as potential markers of metacognitive engagement in revision, and note that these within-text edits are more common in students who had access to the tool. Together, these pilot analyses serve as a first step in testing the tool's theory of action.

1 Introduction

The writing feedback tool, Write-on with Cambi!, was designed to support students in revising their argumentative essays. It highlights key argumentative elements based on annotation guidelines aligned to standards, which have been shown to produce organizational patterns that correlate with rubric scores [1]. These annotations drive the tool's feedback in two primary ways. First, they provide students with a structured overview of their writing. Second, the absence of certain annotations in a student's essay triggers targeted feedback. [3] Beyond annotation-based feedback, the tool flags conventions-related errors (e.g., spelling, punctuation, grammar). It does not auto-correct; instead, it highlights each issue and provides guidance on how to revise it.

The tool is grounded in a theory of action that, at a high level, states: "Students who are guided through a structured review of their essays with immediate, annotated feedback that is well-aligned to teacher instruction will produce essays of higher overall quality."[4] To further theorize the causal mechanism that leads to this outcome, we posit that, by prompting students to examine potentially missing compositional elements and conventions-related errors, the tool elicits metacognitive processes (reviewing, evaluating, and editing) that, in turn, improve essay quality.

This study aims to begin to evaluate this theory of action by the way of the following two key research questions:

1. Do students with access to the tool achieve higher scores across all three dimensions of the scoring rubric? Additionally, how does the effect vary across different teachers in terms of both magnitude and direction?

2. How can we begin to analyze the differences in writing and revision strategies between students who have access to the tool and those who did not? How might we tie this back to the theory of action?

2 Methods

2.1 Randomized Pilot

At the end of the 2024 school year, 11 educators from two states volunteered to pilot the Write on with Cambi! (or Cambi!) tool in their grade 6 through 8th grade classrooms. This pilot study involved 262 seventh grade students within eleven classrooms, with 125 from State A and 137 from State B. During the test, students were randomly assigned access to the tool: 124 did not receive access (control group) and 138 students did receive access to the tool (treatment group).

To begin to assess the impact of the writing tool of Cambi! in student performance, students' essay responses were scored across three dimensions of

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 3: Coordinated Session Papers, pages 115–123
October 27-29, 2025 ©2025 National Council on Measurement in Education (NCME)

the rubric: Conventions, Elaboration and Organization by an automated scoring engine, Autoscore. All students answered the same writing prompt and were scored using the same rubric, which was common across the two states.

We report descriptive statistics (means/SDs), estimate effect sizes (Cohen's d), test group differences (two-sample t-tests; Wilcoxon rank-sum), describe score-point distributions, and examine heterogeneity in treatment effects by teacher/classroom.

2.2 Response Analysis

We collected the full text of each student's essay at 2-minute intervals throughout the writing session, which we will be referring to as "2-minute snapshots" or simply "snapshots."

This process yielded a primary corpus of 4,990 snapshots from 262 unique student participants. Each entry in this corpus contains the student's unique ID, their assigned group (treatment or control), a chronological snapshot sequence number, and the full text of their essay at that moment.

From a qualitative review of two-minute snapshots, we categorized essay revisions into two types: appending—adding new text to the end of the essay—and internal edits—changes made within the previously written text. We further distinguish two forms of internal edits:

1. Surface-level Edits: Minor corrections, oftentimes related to writing conventions, such as spelling, punctuation, and grammar.

2. Substantive Edits: Larger changes or additions within the previously written text of the essay.

Internal edits are of particular interest, as they may signify a deeper level of metacognition, suggesting a shift from automatic drafting to more deliberate and strategic composing.

To analyze revision patterns, we developed a custom algorithm to classify the changes between consecutive 2-minute snapshots. After tokenizing each snapshot's text using the NLTK library, we used a hierarchical classification logic to categorize every change into one of three mutually exclusive types:

1. **Appended Text**: Edits were first checked for location. Any change involving an addition of text at the very end of the previous snapshot was classified as an *Append*.

2. **Surface-level Edits**: If an edit was not an *Append*, its size was evaluated. Any internal change (an insertion, deletion, or replacement within the body of the text) involving 3 words or fewer was classified as a *Surface-Level* Edit.

3. **Substantive Edits**: Any internal edit involving more than 3 words was classified as a Substantive Edit.

This process yielded a count for each of the three edit types for every 2-minute interval. In the charts presented below, the average number of edits is calculated as the total number of edits of a specific type (e.g., surface-level) within a given writing stage, divided by the total number of students in that group.

The algorithms used for this classification of edits is detailed in Algorithm A.1

3 Results

3.1 Randomized Pilot

We analyzed the impact of thewriting tool on student writing by comparing a treatment group (acess to Cambi!) and a control group (without access to Cambi!) across three rubric dimensions: Conventions, Elaboration, and Organization.

3.1.1 Aggregate Results

Across all classrooms, essays written by students with access to Cambi! had higher mean scores on average, as outlined in Table 1. This corresponded to a Cohen's d ranging from 0.25 to 0.26. While this effect size may appear small, it should be noted that a review of over 700 k-12 intervention studies suggest an effect size of over .2 is considered large [2]

To test for statistical significance, we first ran two-sample t-tests, which assume scores are interval data. These tests, as shown in Table 2 confirmed the differences were statistically significant ($p<0.05$). To better account for the ordinal nature of the rubric scores (i.e., the distance between 1 and 2 may not equal the distance between 2 and 3), we also conducted a non-parametric Wilcoxon rank-sum test. The results of this test approached statistical significance at the $p<0.05$ level.

Analysis of the score point distributions revealed specific shifts for the those with access to the writing tool compared to the group without access to the tool, shown in Figure 1

Table 1: Comparison of Mean Scores and Effect Sizes by access to Cambi!

	Mean Score (SD)		
Access to Cambi!	**Conventions**	**Elaboration**	**Organization**
0 (n = 133)	1.50 (.72)	1.21 (.64)	1.46 (.72)
1 (n = 115)	1.67 (.60)	1.37 (.55)	1.63 (.58)
Effect Size (Cohen's d)	.26	.26	.25

Table 2: Two-sample *t*-test Results

Dimension	***t*-Statistic**	***p*-value**
Conventions	-2.056	0.0408
Elaboration	-2.044	0.0420
Organization	-2.016	0.0449

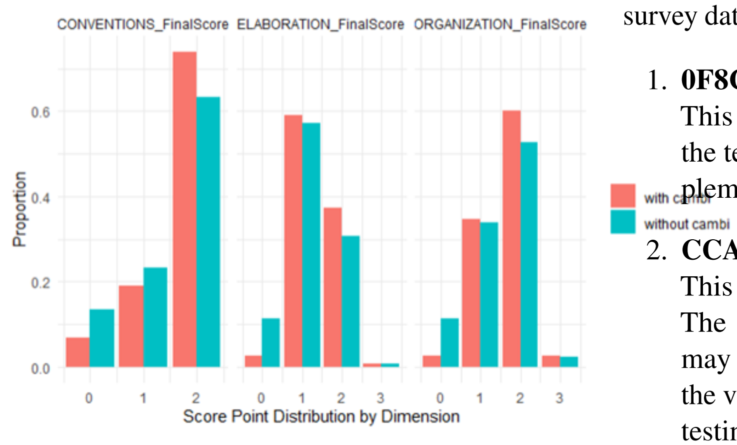

Figure 1: Score Point Distribution by treatment and control group

1. **Conventions**: The Cambi! group received more scores of 2 and fewer scores of 1 and 0.
2. **Elaboration**: The Cambi! group had fewer scores of 0, more scores of 1 and 2, and an equal proportion of 3s.
3. **Organization**: The Cambi! group earned fewer scores of 0, more scores of 1 and 2, and a slightly higher proportion of 3s.

Notably, for Elaboration and Organization, a score of 0 indicated a non-attempt, suggesting the tool helped students overcome initial writing inertia. Additionally, on Elaboration and Organization, no students in either group achieved the maximum score of 4.

3.1.2 Classroom-level Variability

Although aggregate results were positive, the estimated treatment effect varied across classrooms. We take a closer look at one dimension, Organization, to illustrate this variance in Figure 2. For this dimension, 6 of 11 classrooms showed a positive effect for Cambi!, 4 showed a negative effect, and 1 showed no difference.

The variation in results can be illustrated by examining the three largest classrooms, where teacher survey data helps interpret the quantitative findings:

1. **0F8C** *(d=0.32; N: 15 control, 22 treatment)*: This classroom showed a positive effect, but the teacher provided no comment on their implementation strategy.
2. **CCA5** *(d=-0.03; N: 24 control, 14 treatment)*: This classroom showed a negligible effect. The teacher noted that student engagement may have been skewed by low motivation, as the voluntary pilot took place after summative testing at the end of the year.
3. **5F4E** *(d=0.22; N: 29 control, 41 treatment)*: This classroom showed a positive effect. The teacher reported actively scaffolding the tool by going through each feedback tab with students to ensure they understood the suggestions and how to apply them.

In the next section, these pilot results are futher explored to understand how we can begin to analyze the differences in writing and revision strategies between students who have access to the tool and those who did not.

3.2 Response Analysis

In this section, we describe how the revision process—categorized into three edit types—differs between students with and without access to the writing tool.

3.2.1 Overview and Appended Text

An analysis of the overall composition of edits shows that appending new text was a common behavior in both groups. However, as seen in Figure 3, those without access to Cambi! dedicated a

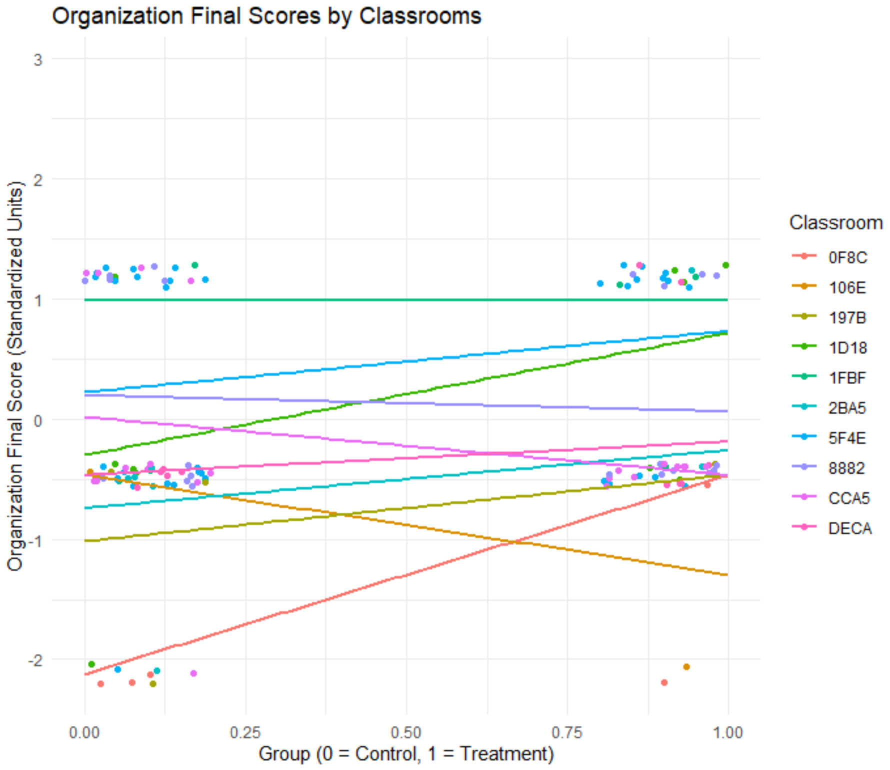

Figure 2: Organization Effect of Access to Cambi! by Teacher/Classroom

larger proportion of their total revision activity to appended edits compared to those who had access to the writing tool.

3.2.2 Surface-level Edits

The timing and frequency of small, surface-level edits revealed a notable difference in writing workflow between the groups.

1. **Overall Trend**: As shown in Figure 4, both groups steadily accumulated surface-level edits throughout the writing session. Notably, while those without access to the writing tool maintained a slightly higher cumulative edit count for the first three quarters, those with access showed a marked acceleration in editing during the final stage (76-100%). This timing aligns with the tool's feedback flow: conventions-related feedback is delivered only after students receive more substantive feedback focused on compositional elements.

2. **Analysis by Score Point**: This trend was most pronounced among students with high scoring essays on the Conventions score. However, the most striking difference was observed among students who ultimately scored a zero on Conventions (Figure 5). Students who had access to the tool but received a lower score showed a high and increasing level of cumulative surface-level edits. In contrast, the control group's essays that received zero scores show almost no cumulative editing activity.

3.2.3 Substantive Edits

The analysis of substantive edits (defined as internal edits involving more than three words) reveals a divergence in revision strategy between the two groups. As shown in Figure 6, students with access to Cambi! consistently accumulated more substantive edits than the control group throughout the entire writing session. The gap between the two groups widened over time, with the treatment group performing a substantially higher number of total substantive revisions by the end of the session.

This difference in behavior was most pronounced among the essays with the highest scores. Figure 7 illustrates the cumulative edits specifically for students who earned a score of 3 on the Organization rubric. For this tier, the treatment group's engagement in substantive revision was higher, with an average of nearly 11 cumulative edits by the end of the writing time. In contrast, their control group peers who also scored a 3 performed very few of these edits, averaging just over 2 by the session's

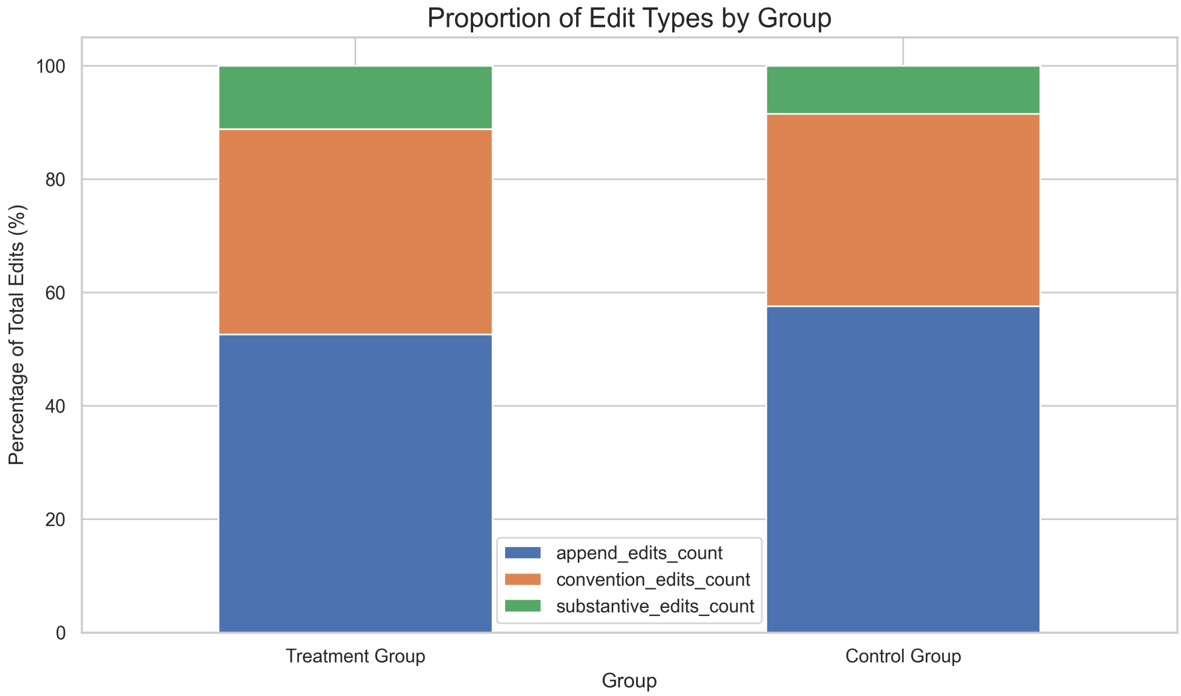

Figure 3: Proportion of Total Edit Types by group.

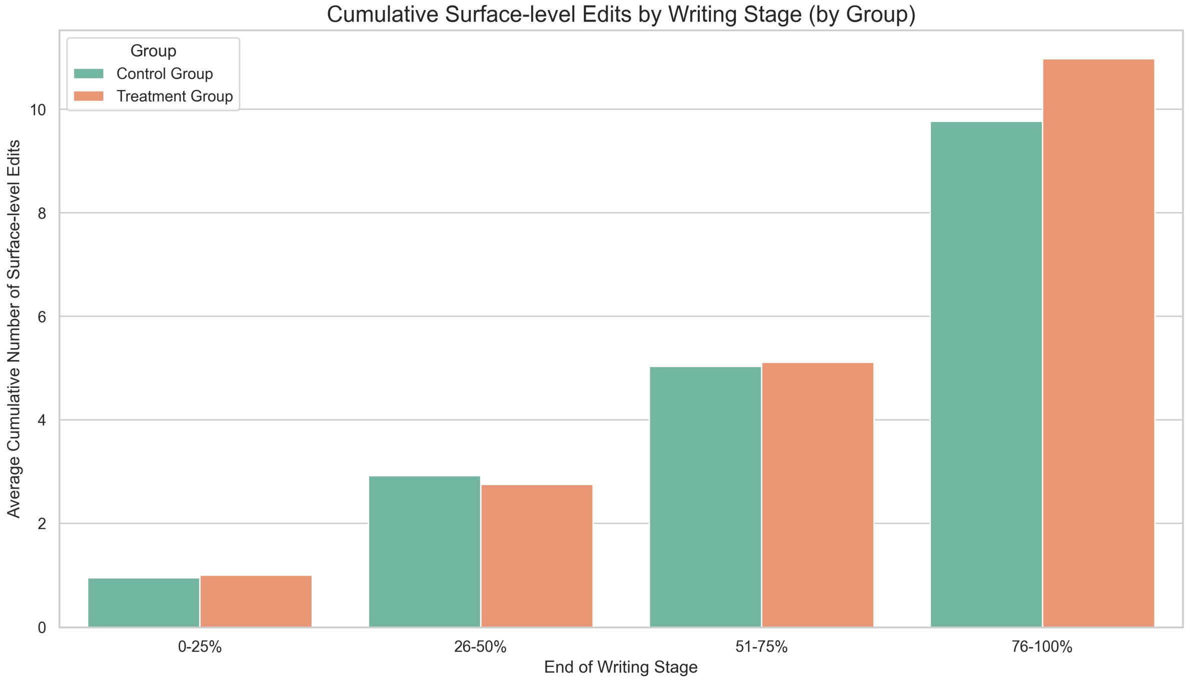

Figure 4: Cumulative Surface-level edits by Writing Stage

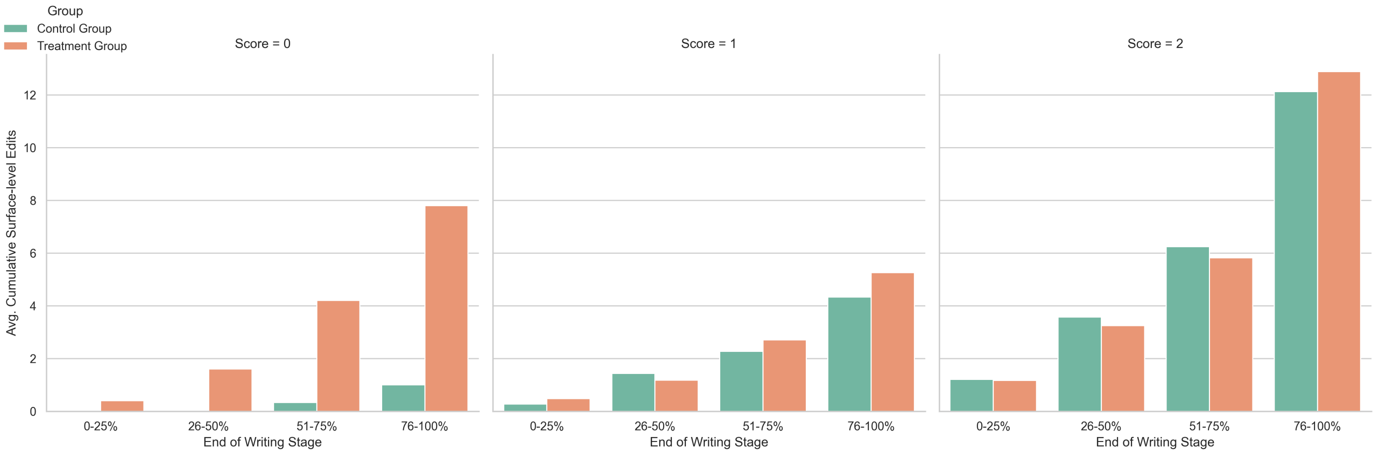

Figure 5: Cumulative Surface-level Edits by Conventions Score Tier

end. The graphs for Organization scores 1 and 2 are in A.2

4 Discussion and Limitations

This work provides preliminary evidence for the effectiveness of an AI-powered writing tool through a pilot study in which students were randomly assigned access. It also introduces methods for exploring the underlying mechanisms that explain how and why the tool influences writing behavior and outcomes, and ties back to the theory of action.

We observed a notable effect of using the tool on student rubric scores, as scored by Autoscore, in the aggregate. Across classes and states, the effect size for each rubric dimension ranged between .25 and .26. While this effect size is large in educational contexts, the outcome of the rubric scores is strongly aligned and scored immediately after the student wrote the essay. As such, with less aligned and further apart outcome variables, we may expect an effect size of a smaller magnitude. Furthermore, observed heterogeneity in the average treatment effect across classrooms, as expected given differences in implementation and baseline writing ability.

We offer several considerations when interpreting the results:

1. **Intent-to-Treat Study**: In this study, we only know if a student was granted access to the tool, but we did not track if the tool was used. The results should be interpreted accordingly. These results provide an estimate of the tool's effectiveness in a real-world setting, where not every student may utilize the tool, they have access to.

2. **Test Fatigue**: The pilot occurred after annual summative assessments, and teachers noted student fatigue. This low-stakes context may have suppressed scores across both groups and masked a larger potential effect.

3. **Control group Behavior**: The control group, aware they lacked access to a new AI tool, may have been less motivated, potentially inflating the observed difference between the groups.

4. **Treatment Diffusion**: Teachers reported helping students interpret Cambi! feedback. It's possible that this guidance was overheard by or shared with control group students, which would weaken the measured effect

In this paper, we also explored methods to begin to analyze differences in writing and revision strategies between students who have access to the tool and those who do not. First, we qualitatively reviewed and categorized the two-minute snapshots into three forms of revisions: appending, surface-level, and substantive. The latter two forms of in-text revisions are aligned with our theory of action that the tool may lead the student to engage in the metacognitive task of reviewing, evaluating, and editing their text.

The findings indicate that students with access to Cambi! tended to shift their efforts from simply appending text toward more internal revisions. This pattern varied across levels of student performance.

For surface-level edits, the tool's impact was most evident among students who received lower scores. As shown in Figure 5, students in the treatment group who ultimately scored a 0 on Conven-

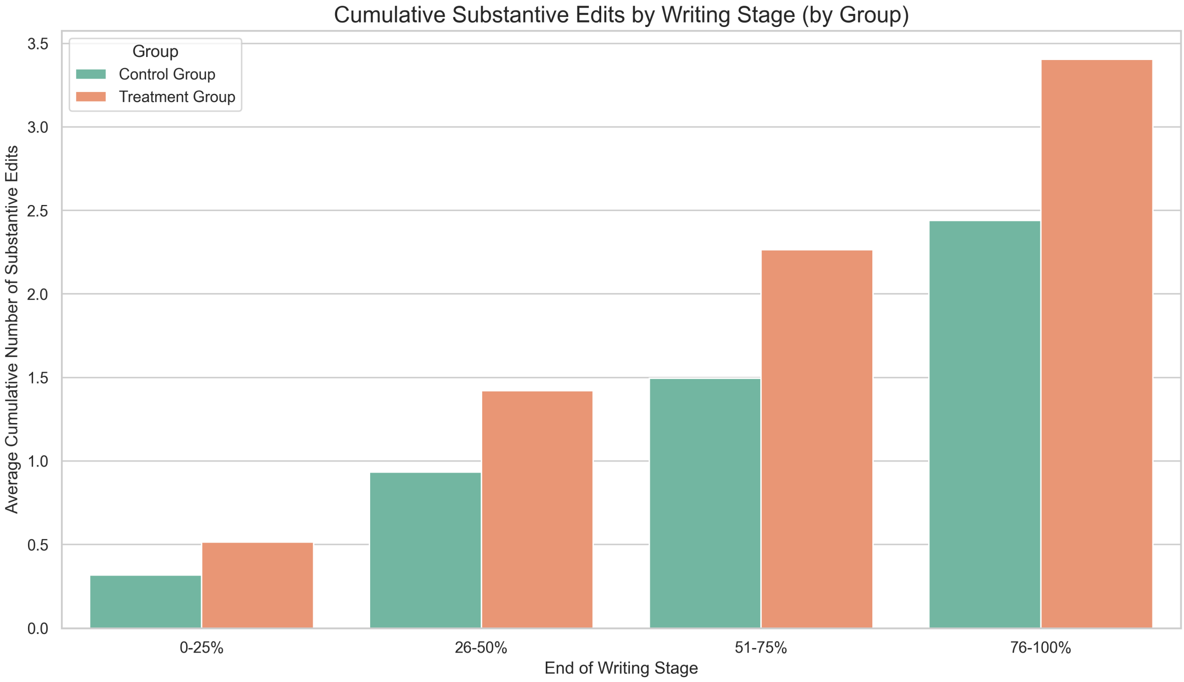

Figure 6: Cumulative Substantive Edits by Writing Stage

tions attempted a notable number of edits, whereas their counterparts in the control group made very few. While these edits did not raise their final scores in this instance, this finding indicates that the tool can prompt engagement from students who might otherwise remain passive.

For substantive edits, the effect was particularly notable among students with higher scores. The data from students who achieved a score of 3 on the Organization rubric shows a substantially higher number of substantive revisions for the treatment group compared to their control group peers (Figure 7). This suggests the tool may act as a scaffold, guiding students who already are capable writers to move beyond surface-level fixes and engage in more complex, structural revision. Future work could also explore different word-count thresholds for differentiating between surface-level and substantive edits.

4.1 Conclusion

Building the evidence base for a writing tool in argumentative writing is ongoing. This paper offers preliminary findings that Write-On with Cambi! can support students and proposes a path for analyzing the mechanisms behind observed score differences. These results serve as a first step in testing the tool's theory of action.

References

[1] Amy Burkhardt, Suhwa Han, Sherri Woolf, Allison Boykin, Frank Rijmen, and Susan Lottridge. 2025. Standards-aligned annotations reveal organizational patterns in argumentative essays at scale. *Frontiers in Education*, 10.

[2] Matthew A. Kraft. 2018. Interpreting effect sizes of education interventions. Technical report, Brown University.

[3] Sue Lottridge, Amy Burkhardt, Christopher Ormerod, Sherri Woolf, Mackenzie Young, Milan Patel, Harry Wang, Julius Frost, Kevin McBeth, Julie Benson, Michael Flynn, Kevin Dwyer, Scott Fitz, Radd Berkheiser, Henry Floyd, Dave Davis, Ben Godek, and Quinell Wilson. 2025. Write on with cambi: The development of an argumentative writing feedback tool. Technical report, Cambium Assessment, Inc.

[4] Sue Lottridge, Chris Ormerod, and Amy Burkhardt. 2025. Development and validation of an AWE system "Write On with Cambi!". In *Proceedings of the National Council on Measurement in Education (NCME)*, Denver, CO.

A Appendix

A.1 Algorithm for edits retrieval

The algorithm used for classifying the edits into three different categories, appended, surface-level and substantive, is outlined in Algorithm 1

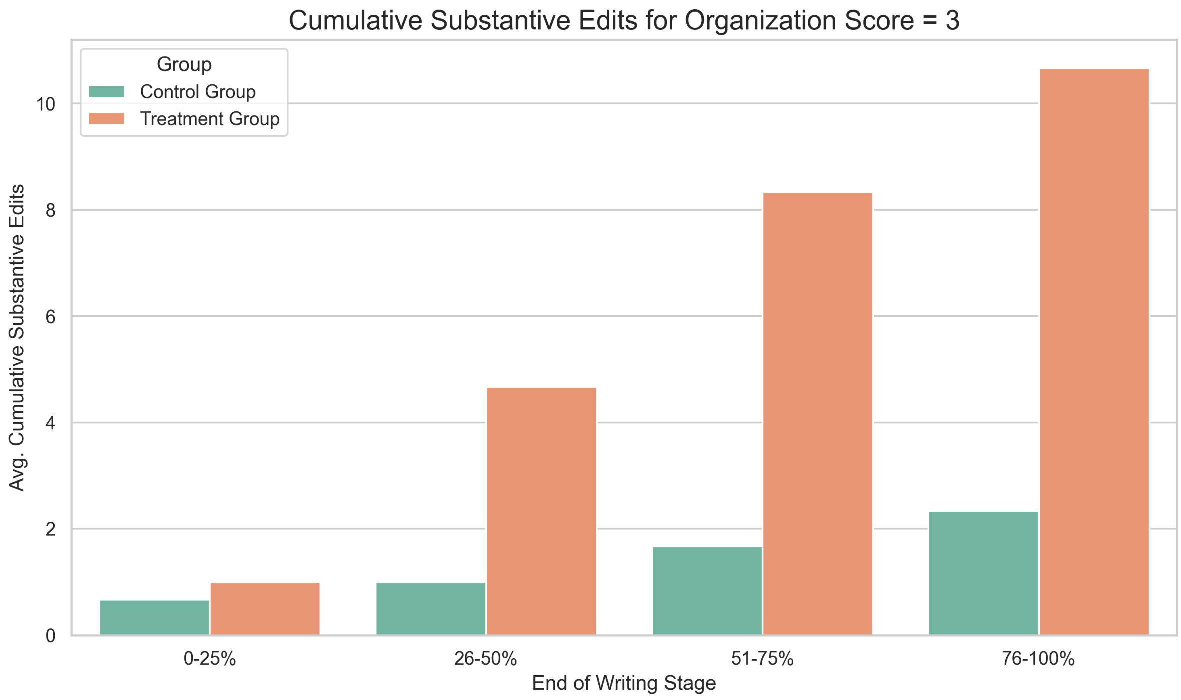

Figure 7: Cumulative Substantive Edits for Students with a Top Organization Score (3)

Algorithm 1 Classification of Revision Edits

```
procedure ClassifyRevision($S_{before}, S_{after}, N_{threshold}$)
    Input: $S_{before}$ (previous text), $S_{after}$ (current text), $N_{threshold}$ (word count limit)
    Output: List of classification labels for each edit
    $W_{before} \leftarrow \text{TokenizeAndClean}(S_{before})$
    $W_{after} \leftarrow \text{TokenizeAndClean}(S_{after})$
    $W_{after} \leftarrow \text{HandleChoppedWord}(W_{before}, W_{after})$
    $Opcodes \leftarrow \text{Diff}(W_{before}, W_{after})$
    $Edits \leftarrow []$
    for all $(tag, i_1, i_2, j_1, j_2) \in Opcodes$ do
        if $tag =$ 'equal' then continue
        end if
        if $tag =$ 'insert' $\wedge i_1 = |W_{before}|$ then
            Append "Appended" to $Edits$
        else if $\max(i_2 - i_1, j_2 - j_1) \leq N_{threshold}$ then
            Append "Surface-level" to $Edits$
        else
            Append "Substantive" to $Edits$
        end if
    end for
    return $Edits$
end procedure
```

A.2 Graphs for Organization scores 1 and 2

Figures 8 and 9 illustrate the cumulative edits by students who earned a score of 1 and 2, respectively, on the Organization rubric

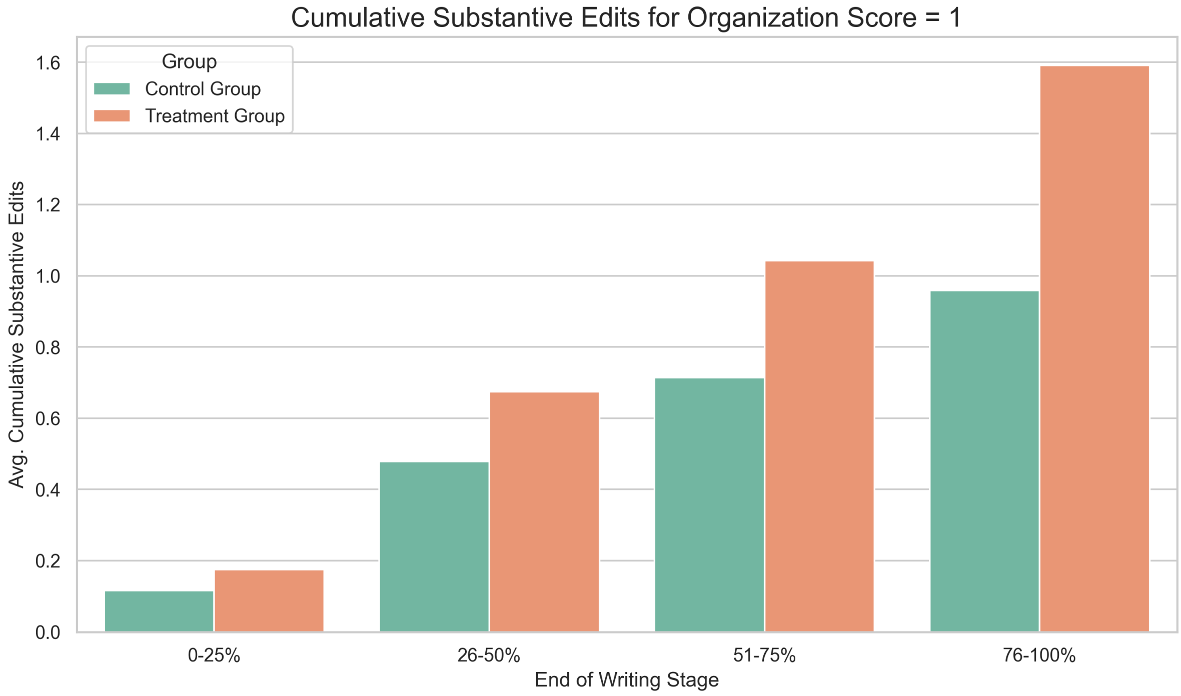

Figure 8: Cumulative Substantive Edits for Students with a Organization Score=1

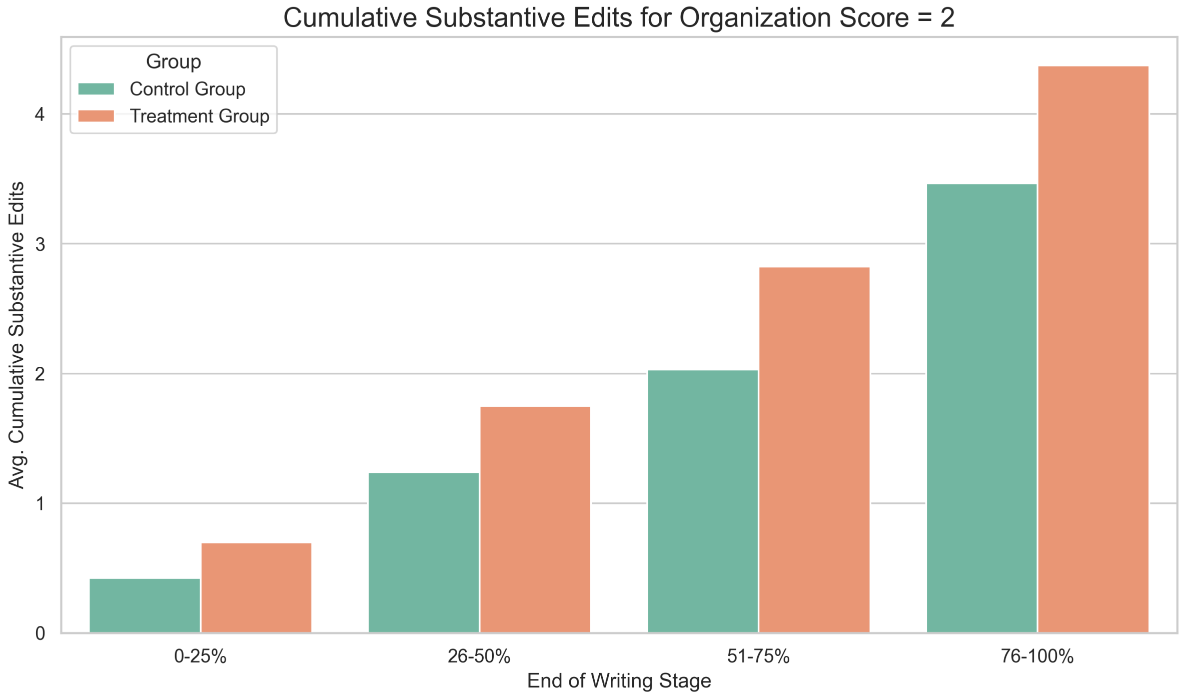

Figure 9: Cumulative Substantive Edits for Students with a Organization Score=2