Automated Essay Scoring in Middle School Writing: Understanding Key Predictors of Students' Growth and Comparing Artificial Intelligence- and Teacher-Generated Scores and Feedback

By Hillary Greene Nolan & Mai Chou Vang
August 2023

Digital Promise®
Accelerating Innovation in Education

Suggested Citation

Greene Nolan, H. L., & Vang, M. C. (2023). *Automated essay scoring in middle school writing: Understanding key predictors of students' growth and comparing artificial intelligence- and teacher-generated scores and feedback*. Digital Promise. doi.org/10.51388/20.500.12265/187; https://digitalpromise.org/wp-content/uploads/2023/09/Topeka-score-analysis.pdf

Acknowledgments

The research team (Hillary Greene Nolan, Mai Chou Vang, and Viki Young) thanks Jess Alanís, Karen Cator, Merijke Coenraad, Briza Diaz, Vanessa Peters Hinton, Megan Pattenhouse, Kristal Brister Philyaw, Teresa Solorzano, Stefani Pautz Stephenson, Jeff Wayman, and Josh Weisgrau for their partnership on Project Topeka.

We are grateful to the many teachers who implemented Project Topeka and shared their experiences through various research activities. In particular, we acknowledge and honor the wisdom and expertise of Kimberly Artis, Rachel Baker, Melissa Castner, Charles Frey, Christy Gibbs, Aida Hadzovic, Elizabeth Hancock, Terry Janssen, Juvy Mojares, Cynthia Orton, Victoria Salcedo, Mary Beth Skerjanec, Millicent Twumasi, Denee Tyler, Adriana Vargas, and Theresa Wampler, who engaged with the Project Topeka team in-depth convenings around instructional AI use, writing instruction, and research findings. The research series amplifies their insights.

This material is based upon work supported by the Bill & Melinda Gates Foundation and Gates Ventures. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Bill & Melinda Gates Foundation or Gates Ventures.

Contact Information

Email: hgreenenolan@digitalpromise.org

Digital Promise:

Washington, DC:
1001 Connecticut Avenue NW, Suite 935
Washington, DC 20036

Redwood City, CA:
702 Marshall Street, Suite 340
Redwood City, CA 94063

Website: https://digitalpromise.org/

Abstract

Providing feedback to students in a sustainable way represents a perennial challenge for secondary teachers of writing. Employing artificial intelligence (AI) tools to give students personalized and immediate feedback holds great promise. Project Topeka offered middle school teachers pre-curated teaching materials, foundational texts and videos, essay prompts, and a platform for students to submit and revise essay drafts with AI-generated scores and feedback. We analyze AI-generated writing scores of 3,233 7th- and 8th-grade students in school year 2021–22 and find that students' growth over time generally was not explained by teachers' (*n*=35) experience or self-reported instructional approaches. We also find that students' growth increased significantly as their baseline score decreased (i.e., a student with the lowest possible baseline grew more than a student with a medium baseline). Lastly, based on an in-person convening of 16 Topeka teachers, we compared their scores and feedback to AI-generated scores and feedback on the same essays, finding that generally the AI tool was more generous, with differences likely driven by teachers' ability to understand the whole essay's success better than the AI tool.

Executive Summary

Project Topeka was launched beginning in 2019 as a collaborative effort among Digital Promise, ThinkCerca, and the Bill & Melinda Gates Foundation with the goal of supporting teachers to give students more opportunities to practice argumentative writing. Facing large class sizes, insufficient planning time, sometimes inadequate preparation to teach writing, and consistently declining writing scores nationally, the project aimed to support teachers in their writing instruction.

Through Project Topeka, teachers could access a platform of instructional materials (e.g., lesson plans, scope & sequence) and assign their students interactive writing assignments (e.g., reading materials, prompt questions). Students wrote and submitted draft essays through the platform, receiving instant artificial intelligence (AI) generated scores on a 4-dimension rubric and dimension-level comments from an automated essay scoring (AES) or automated writing evaluation (AWE) tool on how to improve and resubmit their draft for a better score. Teachers could allow students to revise and resubmit as many times as they wanted, and there were up to six prompt topics to assign.

This research project draws on student score data generated through the use of the Topeka AES tool, teacher survey data, and findings from an in-person convening of a small group of teachers who used Topeka during 2020–21. We investigate the following questions:

1. Which aspects of teacher characteristics, teaching practice, and previous learning predict students' growth on essay attempts over time?
2. How do scores and feedback given by teachers compare to scores and feedback given by AES and AWE?

First, we find that students experienced the most significant writing growth on a final draft when they started at lower baseline (very first draft, no feedback) scores. In other words, students who began at the lowest possible score of 4 overall points out of 16 (i.e., earning a 1 in each dimension) grew significantly more than students who began at, for instance, an 8 (i.e., all 2's on their baseline) or 12 (i.e., all 3's on their baseline). Thus, Topeka participation appeared to benefit lower-performing students than higher-performing students, although, on average, students who began at that lowest level of 4 points still did not meet expectations (12 points) by the final draft—ending, on average, around 8–9 points.

Next, we find that teaching experience and teachers' reports of their typical instructional practices in writing did not significantly predict students' writing growth, with two exceptions. Teachers who believed their students were ready for argumentative writing had students with significant growth on a first prompt, and teachers who felt more confident teaching students how to logically order reasons and evidence had students with significant growth on a first and a second prompt.

Finally, comparing teacher scores to AI-generated scores showed that, generally, teachers were less generous scorers than the AES and had much more nuanced reasoning for their scores than the AES feedback provided to students. The teachers and AES scores showed more agreement at the lowest (1 point) and highest (4 points) scores in each dimension, but whereas the AES was more likely to score students as a 3, the teachers were more likely to score students as a 2. In terms of feedback, teachers

explained that they were able to look for certain indications of what they considered stronger writing that the AES is not yet able to look at, such as voice, personality, and whether the essay made sense holistically rather than line-by-line (the AES's current capability).

It is promising to think that AES tools could offer students more opportunities to practice argumentative writing while at the same time making the planning and grading of teaching more sustainable to teachers, so there is great appeal in improving these tools. Altogether, this study shows the need to factor into the design of AES and other AI tools for writing the knowledge, priorities, and understanding of students that experienced teachers have. Teachers currently using such tools should also be aware of the potential differences and biases in how the AI scores and gives feedback versus how they might. Alongside improving the AI tools, future research should continue to explore the links between teachers' instructional practice around writing and their students' growth; although we found no statistical links here, we believe further research could uncover important relationships between teachers' efforts and students' outcomes to identify high-impact practices teachers could leverage.

Introduction

Assessing student work accurately, efficiently, and in a way that motivates students to improve is a perennial challenge for teachers of writing. Faced with large class sizes and insufficient time to read, respond, and assess extended writing assignments, it is not difficult to see why, nearly two decades ago in 2003, the National Commission on Writing called writing the "neglected R" in K–12 education compared to reading and arithmetic. Additionally, most teachers enter teaching without sufficient experiences as writers themselves or formal preparation to teach writing (Gallagher et al., 2015; Graham, 2019).

Although most teachers recognize the importance of writing to postsecondary success, many report having insufficient time to spend teaching it and grading it, resulting in very little writing being assigned at all (Applebee & Langer, 2011; Graham & Perin, 2007). Perhaps due to having fewer opportunities to practice extended writing tasks, only about one-quarter of United States 8th and 12th graders are deemed "proficient" or "advanced" writers, while the rest are considered "basic" or "below basic" (National Center for Education Statistics, 2012).

Automated essay scoring (AES) and automated writing evaluation (AWE), which rely on artificial intelligence (AI)-based natural language processing to score and give real-time writing-focused feedback to students as they write, have the potential to save teachers time and, perhaps, make it more feasible to give students more writing opportunities. AES has been used extensively in standardized testing contexts, such as college entrance exams. Some research has pointed to the reliability of AES, arguing that AES can be a more objective scorer than a human due to the consistent application of its algorithms (Shermis, 2014). Other studies have shown the limitations of using AES and AWE. A recent study of elementary school writing showed that AES scored writing as accurately as teachers did, except that teachers were better able to distinguish finer-grain differences of students who struggle with writing (bottom 25th percentile) (Chen, Hebert & Wilson, 2022). A study of Chinese undergraduates found that human scorers were able to give English learners more accurate assessments than the AES, and that teachers' feedback grasped the whole essay and elements of style, in addition to adapting feedback for English learners better (Liu & Kunnan, 2016). Another consideration in deploying AES is that, although it can generate ample data for teachers to draw upon in their planning and grading, teachers' beliefs about students from their everyday interactions act as filters for how they view and use assessment data at all (Young & Kim, 2010).

Recognizing this potential, the Bill & Melinda Gates Foundation initiated Project Topeka, an AES and AWE platform targeting 7th and 8th grade students through which teachers could access lesson plans, reading materials for students, and prompt questions, and where students could then submit essay responses, receive real-time scores and feedback, and rewrite and resubmit as often as desired until reaching mastery across four dimensions of writing (Claim & Focus, Support & Evidence, Organization, and Language & Style). This study draws on student writing performance data from Topeka essays paired with teacher scores and feedback, and discussions with 16 teachers who had used Topeka to understand the similarities and differences between human and automated essay scoring and feedback.

In this study, we first explore what predicts students' writing growth on subsequent essay attempts in an effort to see what predicts growth as measured by the AES in particular, and then, we examine differences between human and AES scoring. There we find some similarities and some differences between teacher-generated and AI-generated scores and feedback. Our findings suggest that human and automated essay scoring should ideally be used in complementary ways that draw on the strengths of each—on the one hand, the ability of the AI-based scoring tools to save teachers time, generate informative individual and class-level data, and engage students with immediate feedback teachers cannot so efficiently give and, on the other hand, the ability of the teacher to look at writing holistically, to help students see bigger picture improvements they can make to their writing, and to assess students in a way that factors in both their writing performance and their humanity.

The research questions driving this study are:

- Which aspects of teacher characteristics, teaching practice, and previous learning predict students' growth on essay attempts over time?
- How do scores and feedback given by teachers compare to scores and feedback given by AES and AWE?

Methods

Study Context

Project Topeka was available for any teacher of writing to use in the 2021–22 school year. Digital Promise recruited teachers nationwide, with a focus on middle school Language Arts teachers, to register and use Topeka with their students. Recruitment focused on teachers in middle schools where at least half of students were eligible for free- or reduced-price lunch and identified as students of color. In addition to being able to use Topeka with their students for up to six prompts all year, teachers were also invited to participate in optional research activities—pre- and post-implementation surveys, an interview focused on how they used Topeka resources, and an interview focused on their teaching materials and student work samples.

Throughout 2021–22, staff at Digital Promise sent frequent messages reminding and encouraging teachers to complete prompts with their students. The intended use of Topeka materials involved teachers assigning, first, a baseline prompt to their students, an essay students would complete and submit independently for a baseline score. Then, teachers would assign the same prompt again as a revision prompt, where students could draft, submit, receive real-time, interim scores and feedback, revise, resubmit for additional real-time scores and feedback, and repeat revisions an unlimited number of times until reaching a final score when the teacher officially closed the assignment. Although not all teachers used Topeka in exactly this way (i.e., some did a baseline but no revision, some went straight to revision without a baseline score), most did.

Data and Sample

We created a series of analytic datasets, described next, for this study: an analytic dataset of student writing scores, a subset of that dataset that linked to students' scores their teachers' pre-implementation survey data when available, and a much smaller subset of score data used for an in-person teacher convening, which in turn generated an additional analytic dataset based on teachers' scoring of that smaller subset of essays.

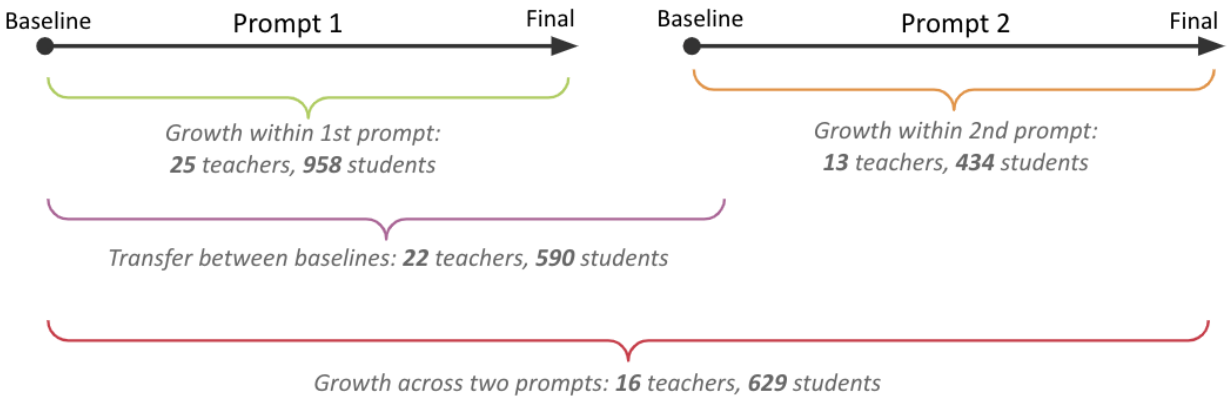
Student writing score data and sample. This study draws from the larger Project Topeka dataset and focuses on school year 2021–22 since school years 2019–20 and 2020–21 were markedly different due to the effects of pandemic-related changes and closures. The Topeka dataset for school year 2021–22 included scores for 7,614 essays from students in grades 5–9; of those, only 420 essays came from 5th-, 6th-, or 9th-graders, so we limited our analysis to 7th- and 8th-graders, who happen to be the target age group for this program.

There were many variations of essays each student could complete. First, Topeka offers six different prompt topics that teachers can assign and/or students can opt to complete (see Appendix Table 1 for prompt topics and details). Therefore, students could complete anywhere from one prompt to all six, and it is possible (though unlikely) they could complete the same prompt more than once. Second, depending on how teachers use Topeka, students can complete (a) just a baseline essay, (b) just a revision essay, or (c) both a baseline essay and a revision essay on the same prompt.

For this study, our final writing score dataset consisted of the writing scores of 3,233 7th- and 8th-grade students, associated with 51 teachers, who had at least one essay from at least one prompt (baseline or revision), and up to four essays from two complete prompts (baseline and revision for two prompts). We did not include students who had attempted a third prompt as only 119 students had, only 13 of whom completed both the baseline and revision for the third prompt.

Corresponding to the four key outcomes in the student score analysis are the four analytic subsamples depicted in Figure 1: (1) growth within a first prompt, consisting of 958 students' first-prompt baseline and revision scores across 25 teachers; (2) growth within a second prompt, consisting of 434 students' second-prompt baseline and revision scores across 13 teachers; (3) transfer or retention across two prompts, consisting of 590 students' first- and second- prompt baseline scores across 22 teachers; and (4) growth across two prompts, consisting of 629 students' first-prompt baseline and second-prompt revision scores across 16 teachers.

Figure 1. Student Score Analytic Samples



Teacher pre-implementation survey data and sample. As part of the 2021–22 Project Topeka research study, 246 teachers completed a pre-implementation survey about their pre-Topeka experiences and perceptions around teaching, writing, and their students as teachers engaged in writing instruction. To explore relationships between teachers' experiences and perceptions and students' writing performance, we merged teachers' survey responses from the pre-implementation survey onto our writing score dataset. Of the 51 teachers in our writing score dataset, 35 teachers had completed a pre-implementation survey. Therefore, our analysis of teacher experiences and perceptions in relation to student writing performance draws on the surveys of 35 teachers and the scores of 2,308 associated students.

Teacher convening scoring data and sample. We held an in-person convening with 16 Topeka teachers in June 2022 during which teachers scored a set of 10 essays each. We then compare their scores to the AES scores students had received.

Essays were selected for this event based on a number of criteria. First, we did not include essays from students of any of the attending teachers. We included essays completed in school years 2020–21 and 2021–22, restricted to teachers who had completed one full prompt with their class (baseline and final). We further restricted potential essays only to teachers who had completed a pre-implementation survey (in their respective implementation year) in order for us to use teacher-reported demographics of their students. We only included teachers fitting the above criteria whose free- or reduced-price lunch eligible population was at least 51 percent.

Those criteria yielded essays across 21 teachers. From there, we selected essays that balanced 7th- and 8th-graders, prompt topics, and score distributions. In the end, we selected 170 essays across 21 teachers based on these criteria; we used 160 at the convening since 16 teachers attended and each scored 10. Appendix Table 2 displays information about the 170 essays.

Of the 16 teachers who attended the in-person June convening, 10 attended the virtual November convening, where we elicited their reactions to the similarities and differences between human and AES

scores and feedback, echoing discussions we had already facilitated at the June convening but giving us a way to see if their reactions were consistent over time.

Measures

Student writing score measures. For each essay, a student could receive a maximum of eight scores from the AES: one baseline score and one revised score for each of four dimensions. Dimensions included Support & Evidence, Claim & Focus, Organization, and Language & Style, and these were scored on a scale of 1–4, with 4 being the top score. For our analysis, we also created an overall score for each essay based on summing the four dimension scores, so students could have an overall score of 4–16, with 16 being the top score.

Additionally, we calculated changes, when possible, between students' baseline and revision scores corresponding to our four key outcomes represented above in Figure 1 and below in Table 1. We calculated these at the overall and dimension levels. Table 1 shows the average baseline and revision scores for each prompt, as well as average differences for the four key outcome scores: change within the first prompt, change within the second prompt, change across two prompts (from first baseline to second revision), and transfer across two prompts (from first baseline to second baseline).

Table 1. Average Baseline, Revision, Change, and Transfer Scores Across Two Prompts

	Prompt 1			Prompt 2			Across-Prompt Transfer (Base 1 to Base 2) (Outcome 3)	Across-Prompt Change (Base 1 to Final 2) (Outcome 4)
	Baseline B	Final	Change (Outcome 1)	Baseline	Final	Change (Outcome 2)		
Support & Evidence	1.98	2.57	+0.56	2.20	2.55	+0.50	+0.11	+0.42
Claim & Focus	2.03	2.65	+0.57	2.30	2.63	+0.50	+0.15	+0.50
Organization	2.18	2.80	+0.55	2.51	2.84	+0.45	+0.26	+0.56
Language & Style	2.20	2.81	+0.53	2.47	2.81	+0.49	+0.17	+0.50
Sum Score	8.39	10.82	+2.21	9.48	10.82	+1.93	+0.67	+1.98
n Students	2,435	1,756	958	672	799	434	590	629
n Teachers	42	31	25	21	15	13	22	16

We explored whether scores differed according to the topic of the prompt, of which there were six, and based on whether the student was a 7th- or 8th-grader. We found some evidence that prompt topic and grade level did influence scores, and as we explain below, we included these as controls in our analytic models; see Appendix Table 3 for average scores by prompt topic and grade level.

Teacher pre-implementation survey measures. Where teacher survey measures were used in this analysis, we used information directly as reported on the survey (e.g., teachers' reported years of experience as a continuous variable) or created dichotomous versions for ease of interpretation (e.g., item about ranges of revision opportunities teachers offered students collapsed to more than two versus fewer than two). The latter decisions were based on examining the distributions of the responses and also being conceptually clear. Table 2 shows descriptive information for the survey items used in this analysis.

Table 2. Teacher Pre-Implementation Survey Responses on Items Used as Predictors in Analysis

Item	Mean (sd) / %
Years teaching	17.41 (8.61)
Years teaching Language Arts	15.79 (7.65)
Offer 2 or more revisions on typical assignment (vs. 0–1)	98.3%
Give feedback monthly or more (vs. less frequently)	44.5%
Writing process taught monthly or more (vs. less frequently)	28.9%
Extended writing opportunities monthly or more (vs. less frequently)	11.1%
Agree/strongly agree my students enjoy writing*	22.7%
Agree/strongly agree my students are ready for argumentative writing*	25.7%
Agree/strongly agree I feel prepared to teach argumentative writing*	71.0%
Confidence in teaching the ordering of reasons/evidence in a logical way (1–4)	3.37 (0.73)

Note: n=35. * vs. disagree or strongly disagree.

Analytic Method

Student score analytic method. We used 2-level hierarchical linear models with students nested in teachers to estimate students' average growth on our four key outcome measures as a function of students' corresponding baseline scores. Therefore, for our first outcome—growth within the first prompt—we estimated the change between students' baseline and final first prompt scores as a function of their first prompt baseline score. For our second key outcome—growth within the second prompt—we repeated the first model, substituting change between students' baseline and final second prompt scores as the outcome and second prompt baseline as the predictor. For our third outcome—

transfer from the first baseline to the second baseline —we modeled change from first to second baseline as a function of the first baseline score. And for our fourth outcome—growth from the first baseline to the second final—we modeled change from first baseline to second final as a function of first baseline.

In all models, we treated our outcome measures as continuous and our predictor measures (baseline scores) as ordinal in order to see the predicted growth at each starting score. Throughout, we also included as controls students' grade level and the prompt topic (based on notable differences in scores across those groups), and the part of the school year during which the essay was completed (before holiday break vs. after holiday break) to account for students' natural progress within a school year.

Pre-implementation survey analytic method. To analyze relationships between teachers' experiences and perceptions and their students' writing performance, we used the same modeling approach as with the student scores, making the predictors be our focal teacher survey items. We still included the appropriate baseline score as a control in each model, but in order to avoid dropping the sample and since we found that the estimates were similar with and without controlling for grade level, prompt topic, and season, we did not include them as controls in the results shared here.

Teacher convening analytic method. For each of the 160 papers, we created a sum score for teachers by adding the teacher scores of the four domains. We also created a sum score for the AI scoring tool. To compare teacher scores and AES scores across all papers, we conducted a paired t-test. We also used the same test to see if there were significant differences in scores between teachers and the AES by dimension.

Findings

Analysis of student writing scores suggests Topeka benefitted most the lowest-scoring students at baseline. For each of our four focal outcomes—growth within the first and second prompts, transfer from the first to second prompt baseline, and growth from the first baseline to the second final—students with lower baseline scores experienced, on average, the highest growth, whereas students who started with mid-range or higher scores were predicted to have limited or no significant change across essay attempts. Table 3 shows estimates from these models. Also, the first two columns of Table 3 show the number of students who started out in each baseline at each possible sum score. Within each band—the white band of students with the lowest starting scores, the light gray band of students with the middle starting scores, and the darker gray band of students with the highest starting scores—we see that more students begin at a score of 4, 8, and 12, or roughly all 1's, all 2's, and all 3's.

Table 3. Growth on Each Outcome's Overall Score Predicted by Corresponding Baseline Overall Score

	n at 1st Baseline (*n* = 1,535)	*n* at 2nd Baseline (*n* = 965)	Outcome 1: Growth From 1st Baseline to 1st Final (*n* = 958)	Outcome 2: Growth From 2nd Baseline to 2nd Final (*n* = 434)	Outcome 3: Transfer from 1st Baseline to 2nd Baseline (*n* = 590)	Outcome 4: Growth from 1st Baseline to 2nd Final (*n* = 629)
Baseline score (16 = reference group)						
4	513	139	3.88 (0.54)***	4.11 (0.84)***	5.41 (0.78)***	5.96 (0.76)***
5	114	38	3.23 (0.61)***	6.19 (0.96)***	5.65 (0.86)***	5.20 (0.85)***
6	90	25	4.46 (0.65)***	4.24 (0.96)***	5.38 (0.97)***	4.82 (0.90)***
7	111	37	2.26 (0.61)***	2.91 (0.90)**	4.93 (0.92)***	4.50 (0.86)***
8	570	190	2.69 (0.54)***	2.57 (0.81)**	3.25 (0.77)***	4.22 (0.74)***
9	161	56	2.59 (0.56)***	2.86 (0.83)**	3.12 (0.81)***	3.67 (0.77)***
10	184	101	2.06 (0.55)***	2.09 (0.83)*	2.69 (0.80)**	3.75 (0.77)***
11	130	75	1.53 (0.56)**	1.87 (0.84)*	2.32 (0.83)**	2.74 (0.79)**
12	337	186	1.44 (0.53)**	1.59 (0.80)*	1.43 (0.76)	2.18 (0.74)**
13	89	37	1.53 (0.59)**	1.47 (0.94)	1.20 (0.83)	2.28 (0.79)**
14	67	33	1.00 (0.59)	1.30 (0.96)	0.50 (0.86)	0.85 (0.86)
15	36	25	0.66 (0.68)	0.34 (0.98)	-0.18 (0.90)	1.45 (0.99)

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Across outcomes, our findings suggest students who start at the lowest baseline scores stand to gain the most from the Topeka process. For example, a student whose first baseline is the lowest score of 4, meaning they scored a 1 in each dimension, on average, rises to 8 points by the final on that prompt, to over 9 points by the start of the second prompt, and to 10 points by the end of the second prompt. Within the first and second prompts, similar gains are evident for students whose initial baselines are 5 or 6 points as well. Looking at growth across two prompts, from first to second baseline or from first baseline to second final, students who start with these lower overall scores of 4–7, meaning they earned

mostly or partly 1's on their first try, gain, on average, around 5 points, rising considerably to scores of 9–12.

In the middle band of baseline scores are those earning mostly 2's on their baselines, with overall scores of 8–11 initially. Our analysis suggests these students still experience significant growth within prompts and across two prompts, though the magnitude is somewhat less than those who begin in the lowest band. Middle-band baseline scorers gain, on average, 2–3 points within prompts and 3–4 points across prompts, which could help them rise to that higher band.

In the highest band of baseline scores are those starting with scores of 12 or up, pointing to dimension-level scores of 3's and 4's. At this level, with perhaps less room to improve at all on the scoring rubric, we see fewer instances of significant change. Students who begin at 12, possibly earning all 3's, are still predicted to have about a 1–2 point increase by the end of the first prompt, the end of the second prompt, and across both prompts. For students starting at or above 14 points, it appears much more difficult to experience a significant increase in subsequent essay scores.

Generally, we see the same patterns at the dimension level, as shown in Appendix Table 4. Students whose initial baseline score is a 1 in any dimension show significantly higher magnitude gains on subsequent essays than students who begin at a 2; likewise, students who begin with 3's have lower magnitude gains than students who begin with 2's.

Some teacher-reported perceptions predicted students' writing performance, but amount of teaching experience and most teacher-reported instructional practices did not. Expecting that some aspects of teaching might influence students' writing performance, we next modeled our four student writing performance outcomes as a function of teachers' perceptions and reports. On the whole, we found little evidence that students' writing performance is related to what their teachers reported about their experience, their instructional practices, and their perceptions of their teaching and of their students' learning, though a few pieces stood out. All results are detailed in Table 4; results were very similar at the dimension level, so only results using overall scores are shown.

Table 4. Growth on Key Outcome Overall Score as Predicted by Teacher Survey Information

	Outcome 1: Growth From First Baseline to First Final (*n* = 673)	Outcome 2: Growth From Second Baseline to Second Final (*n* = 270)	Outcome 3: Transfer from First Baseline to Second Baseline (*n* = 422)	Outcome 4: Growth from First Baseline to Second Final (*n* = 304)
Years teaching	0.02 (0.03)	0.03 (0.04)	0.07 (0.04)	0.02 (0.04)
Years teaching Language Arts	-0.00 (0.04)	0.05 (0.05)	0.06 (0.05)	0.05 (0.05)
Offer 2 or more revisions on typical assignment (vs. 0–1)	0.35 (0.59)	-0.83 (0.66)	-1.07 (0.76)	-0.22 (0.72)
Give feedback monthly or more (vs. less frequently)	-0.07 (0.60)	-0.73 (0.68)	-1.35 (0.70)	0.06 (0.70)
Writing process taught monthly or more (vs. less)	-0.02 (0.68)	-0.41 (0.83)	0.77 (0.82)	0.28 (0.73)
Extended writing monthly or more (vs. less frequently)	-1.72 (1.22)	NA[a]		2.10 (0.77)**
Agree/strongly agree my students enjoy writing (vs. disagree/strongly disagree)	1.11 (0.75)	-0.44 (0.82)	1.26 (0.83)	-0.27 (0.78)
Agree/strongly agree my students are ready for argumentative writing	1.30 (0.64)*	0.84 (0.80)	1.04 (0.79)	0.16 (0.79)
Agree/strongly agree I'm prepared to teach argumentative writing	0.95 (0.57)	1.16 (0.62)	0.85 (0.92)	0.14 (0.81)
Confidence in teaching ordering reasons and evidence in a logical manner (scale of 1-4)	0.99 (0.39)*	1.22 (0.36)**	0.69 (0.51)	0.01 (0.52)

Notes: [a] We could not model these two outcomes for the predictor of whether teachers reported offering monthly extended writing opportunities because, for teachers with survey responses to that question and available baseline and revision data on outcomes 2 and 3, there was no variation in the survey responses.

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

First, students' growth within the first and the second prompt was strongly predicted by teachers' self-reported confidence in one particular instructional skill: teaching students how to order reasons and evidence in logical ways in their writing. Teachers who felt more confident in this part of their instruction had students who grew significantly from the first baseline to the first final *and* from the second baseline to the second final—with the magnitude of growth slightly higher on the second prompt. Confidence in this area, however, was not significantly associated with transfer or growth across prompts, which merits further research.

Another notable finding was the significant relationship between teachers who believed their students were well-prepared for argumentative writing and students experiencing writing growth. This relationship was significant within the first prompt and trending positive within the second prompt, likely approaching significance if the sample were higher.

Students' growth in writing was unrelated in our analyses to teachers' years of experience, to their reported instructional practices, to how prepared they felt for teaching writing, and to how much they thought their students enjoyed writing. Although not significant, our results showed growth trending negative when teachers reported engaging in more frequent writing opportunities, more frequent teaching of the writing process, more frequent feedback, and more frequent revisions.

Compared to human teacher scorers (M = 7.61, SD = 2.64), the AES tool (M = 8.80, SD = 3.20) appeared to be more generous in scoring students overall, t(159) = -4.17, p < .001. This held true also at the domain level. There were statistically significant differences in AES and teacher scores for three of the four dimensions: Support & Evidence, Organization, and Language & Style, where the AES tool scored higher than teachers did (Table 5). In the Claim & Focus domain, we observed the same trend, and the difference might have been statistically significant with a larger sample.

Table 5. Teacher and AI Mean Scores by Domain and Overall

	Teacher Scores	AI Scores	
	Mean (*sd*)	**Mean (*sd*)**	**Mean difference**
Support & Evidence	1.87 (0.71)	2.04 (0.79)	-0.18*
Claim & Focus	2.07 (0.81)	2.18 (0.86)	-0.11
Organization	1.88 (0.80)	2.26 (0.88)	-0.38***
Language & Style	1.79 (0.74)	2.33 (0.87)	-0.53***
Overall	7.61 (2.64)	8.80 (3.20)	-1.19***

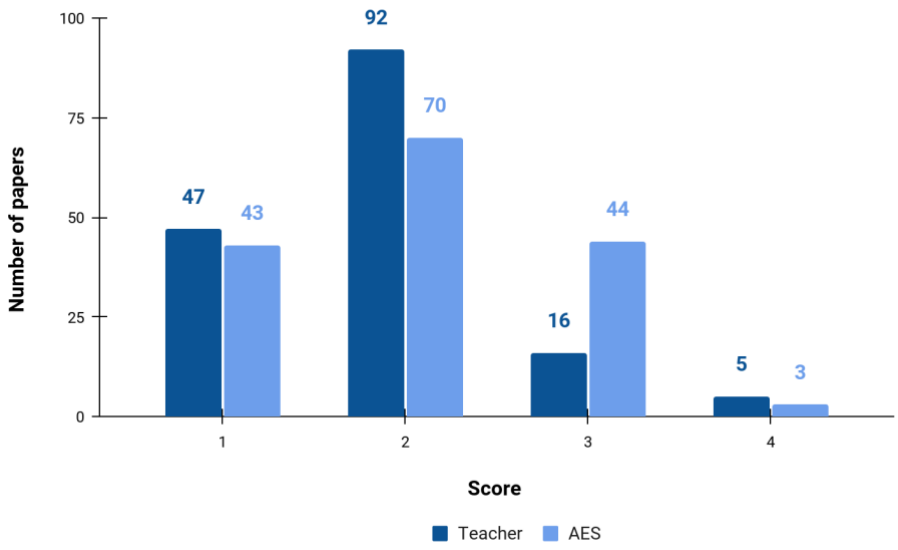
Notes: Dimension scores range from 1–4; overall scores from 4–16. *n*=160 essays.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Figures 2–5 illustrate domain-level differences between teacher and AES scores for all 160 essays that were scored by our convened group of teachers. Additionally, in our June convening with teachers after

they scored the essays, we presented Figures 2–5 to them and listened as they reacted and grappled with the similarities and differences, sharing their reasoning for their own scores and their theories on the AES scores. At our November convening, we presented these graphs to teachers again to gather additional reactions about the similarities and differences between their scores and the AES scores.

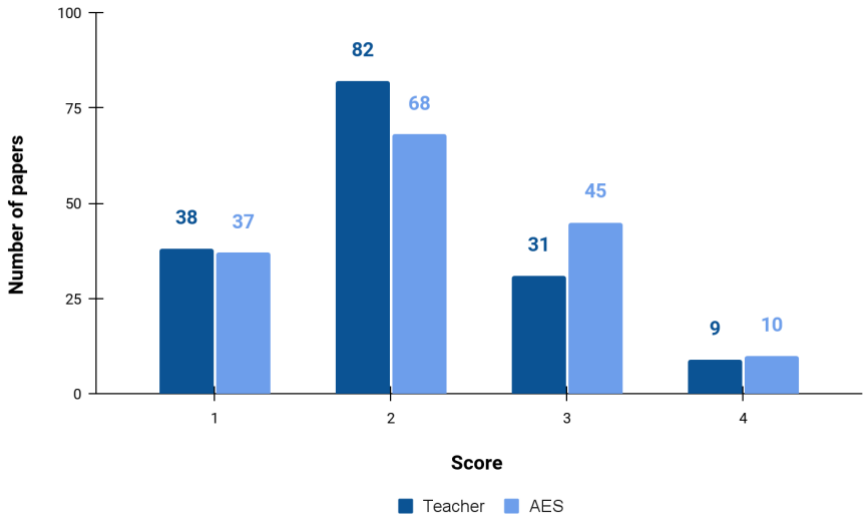
First, in the domain of Support & Evidence (Figure 2), the majority of teachers scored students at a 2, whereas the AES gave a greater range of scores from 1–3. There was, however, solid agreement at scores of 1 and 4 between the teachers and the AES. In this domain, teachers reported scoring lower than the AI tool because they didn't consider simple summaries of sources in the essays to constitute evidence and instead looked for specific sources that supported and connected back to the claim.

Figure 2. Comparing AI- and Teacher-Generated Support & Evidence Scores



In the domain of Claim & Focus (Figure 3), we again saw similar agreement between the teachers and AES at the lowest and highest scores and less agreement at the middle scores of 2 and 3. Just like in the domain of Support & Evidence, teachers were more likely to give a 2 as a score, whereas the AES appeared more likely to give a 3 as a score, suggesting teachers err on the side, when deliberating between a 2 (approaching expectations) and 3 (meeting expectations), of indicating that a student is still on their way and not "there" yet. In discussions, teachers reported some disagreement with what they suspected the AES emphasized in its scoring. For example, some teachers shared that the AES tool tagged a sentence as the claim for the essay, which they disagreed was the correct claim. In other instances, while the AES tool tagged the correct sentence as the claim in the essay, teachers disagreed with where the claim was located. They thought the tool should have provided feedback on where the claim should be moved in the essay rather than simply highlighting that there was one.

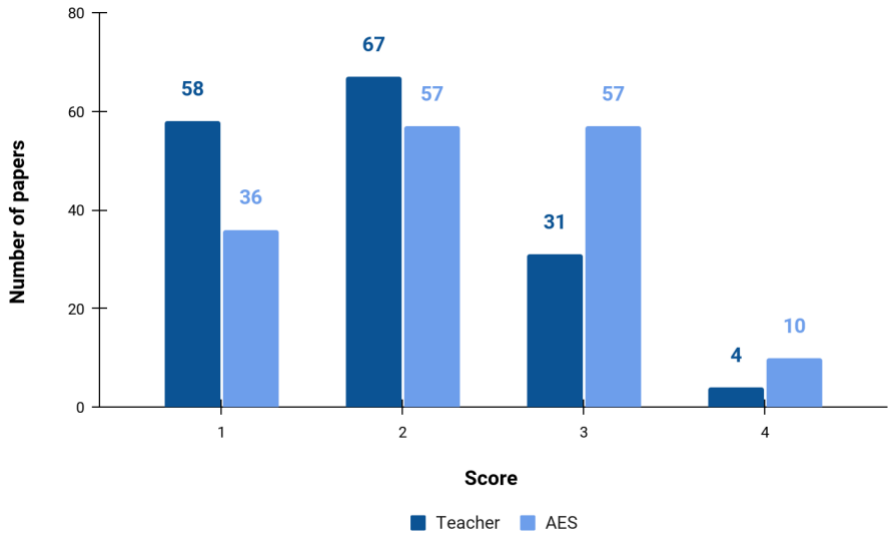
Figure 3. Comparing AI- and Teacher-Generated Claim & Focus Scores



In the domain of Organization (Figure 4), we see a different trend than we did with Support & Evidence and Claim & Focus. Instead of agreeing at the tails and disagreeing in the middle as with those two domains, on Organization, we find that teachers skew toward the low end, with the majority of teachers giving a 1 or 2 score, whereas AES exhibits a normal distribution, with the majority giving a 2 or 3 score.

Teachers' tendency to consider essays as 1's or 2's in Organization is likely due to the ability of a human scorer to see and grasp essays on a much more holistic level than AES systems can currently do. To date, AES tools read and provide feedback line by line. Because feedback is given sentence by sentence, teachers in our discussions worried that students do not necessarily register that feedback should be applied to the whole essay. As a result, any edits students make are sentence by sentence, which teachers reported can lead to an incoherent essay without a logical flow. For example, one teacher shared that in one essay, the AES tool highlighted transitions, but was not able to detect that the ideas between the transitions did not connect. Teachers suspected that the AES tool looks for key or transition words in the text, and then scores the essay as having met the Organization domain 3 or 4 score. However, teachers saw such transitional words in the absence of connected thoughts as insufficient and thus gave lower scores for that domain.

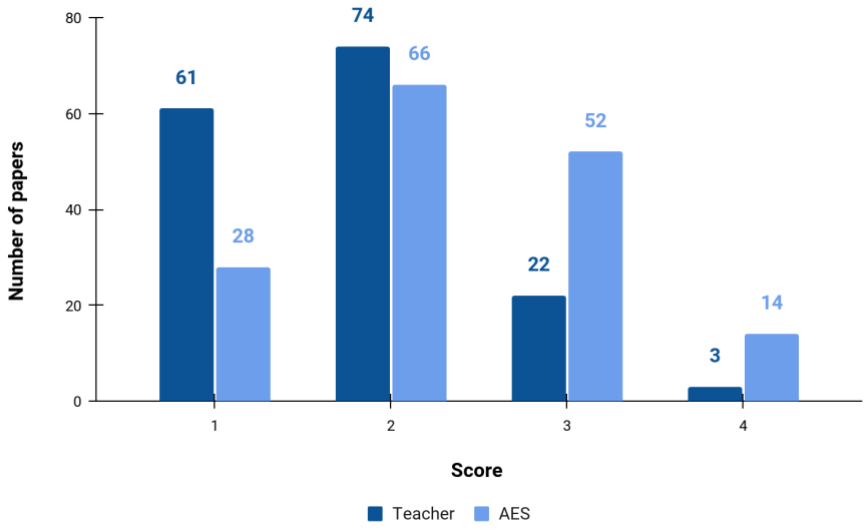
Figure 4. Comparing AI- and Teacher-Generated Organization Scores



Finally, in the dimension of Language & Style (Figure 5), a similar pattern to the domain of Organization emerged, with the vast majority of teachers assessing the essays at a 1 or 2, while the majority of AES scores were scored a 2 or 3. When asked to reflect, teachers said that in this domain, they looked for evidence of student voice and personality and could often detect it, scoring them higher in this domain when their voice came through even if they had challenges in other domains. A couple of teachers questioned if AES could really detect and judge Language & Style, especially style which they considered as student voice.

In many cases, however, teachers placed more emphasis on grammatical conventions, seeing misspellings and run-on sentences as reasons for a lower score. Some suspected that the AES might consider multiple errors as one total error if a student repeatedly, for instance, used run-on sentences, whereas the teachers felt that making the same error repeatedly warranted a lower score. AES did not tag incomplete or run-on sentences. For instance, AES would provide positive feedback on having varying sentence structure for passages even though they solely included author and title, which teachers could easily detect as incomplete or insufficient sentences, but which AES detected as formal, sophisticated language—indeed because it was not student language. Also, teachers wanted students to use formal language in their writing, so phrases such as "I think" and "I believe" did not warrant a high score in their view.

Figure 5. Comparing AI- and Teacher-Generated Language & Style Scores



The reasoning behind teachers' scores was deeper and more nuanced than the feedback given by the AES tool. Teachers generally observed that the AES tool was more generous or lenient in its feedback and scores. AES feedback is given at the sentence level, resulting in students making edits that do not necessarily make the essay better or coherent as a whole. Teachers suspected that the AES tool also scores based on key words whereas they read and score essays more holistically. One teacher thought that the AES scores appeared to be "true bell curves" whereas teacher scores skewed more toward the lower end, though our sense is that both AES and teacher scores skew lower. Additionally, the teachers understood that the AES was operating on a set of criteria, and some shared that they had tried or had

students try to test and figure out what those criteria were, and how to conquer or trick the AES. In other words, rather than using this AES and coming to believe that it assessed writing in an accurate and sophisticated way, several teachers realized its limitations and pivoted to teaching students about the way a system like this works and how to assess if the AES feedback makes sense or is appropriate before addressing it.

Making Sense of the Findings and Looking Ahead

Assessing student learning is one of the many complex parts of teaching. Assessing a student's understanding involves the need to be accurate, while also being sensitive to how such an assessment might motivate or elicit an emotional response and factoring in how an assessment will communicate messages about that student to potential audiences (e.g., the student, a parent, an administrator). It was clear in our discussions with Topeka teachers that, in giving students scores and feedback, they factor in a complex web of information about their specific students, their general expectations for students, and their understanding of good writing. They are able to see many factors the AES would not be able to see, from the style of an essay to the state of a student, and their comments showed that they take this expertise seriously. In this way, areas that were less visible to the AI system were actually a focal point for teachers in assessing student work. In one sense, our study shows that any AI-based platform has a lot of catching up to do in order to be able to capture the complexity and nuance teachers capture in their scoring and feedback on writing.

Nevertheless, there is good reason to catch AES tools up. Transferring some of teachers' assessment responsibilities over to an AI-based system could introduce objectivity, save teachers time, and remove some of the emotional complexity of assessment, which could sometimes interfere with the accuracy. In fact, in interviews we conducted with Topeka teachers in another part of this research, some teachers shared that their students seemed more willing to accept feedback from the AI tool because they perceived that feedback as more objective than the teacher's feedback. Additionally, with many teachers lacking sufficient understanding of or confidence in writing themselves, and with most never having completed teacher preparation or professional development on the teaching of writing, AI tools could fill in knowledge and pedagogy gaps some teachers might have. Indeed, one teacher we interviewed felt using this platform finally gave her a set of writing vocabulary she never herself had that she could now use to build a teaching practice around writing. In this way, AI could be not just a time-saving for teachers with expanded learning opportunities for students; done well, it could also be an embedded professional development for teachers.

Additionally, one area of promise is that students engaged in Topeka who began at the lowest levels of writing performance did exhibit tangible growth in subsequent writing attempts, showing the promise this extended writing practice could hold for the most vulnerable students. Keeping in mind that recent research has shown teachers are better at assessing the writing specifically of struggling writers, it seems that giving these students more writing opportunities and engaging them through a game-like revise and resubmit platform, when paired with heavily involved teachers who also score and give feedback, could particularly help struggling writers.

Our finding that student writing growth did not appear statistically related to anything teachers reported about their own experiences, their status quo writing instruction, and their beliefs about teaching writing merits further exploration. It is possible that facets of teachers' beliefs and instruction that we did not ask about actually would predict student growth. It is also possible that mediating effects are at play and that if, for example, we asked students about their perceptions of their teacher's instruction, we would find stronger relationships. Teachers' self-report data about their beliefs and instruction could be inaccurate. It could also be that student writing growth is less strongly tied to instruction at all than we might have expected and that, instead, students are learning (or not learning) to write based on factors outside of the classroom.

The finding that teachers' confidence in their ability to teach the more sophisticated skill of ordering reasons and evidence in a logical manner suggests that perhaps a key to explore as a lever in achieving more student growth is identifying teachers who are more well-developed in their own writing skills and who grasp the logical structures of writing. Though previous research suggests most teachers are not experienced in writing or trained in the teaching of writing, it is likely that some are, and those teachers' practices and their students' growth are worth investigating further.

References

Applebee, A., & Langer, J. (2011). A snapshot of writing instruction in middle schools and high schools. *English Journal*, *100*(6), 14–27. https://www.jstor.org/stable/23047875

Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *American Educational Research Journal,* 59(6), 1122–1156. https://doi.org/10.3102/00028312221106773

Gallagher, H. A., Woodworth, K. R., & Arshan, N. L. (2015). *Impact of the National Writing Project's College-Ready Writers Program on teachers and students.* SRI International. https://www.sri.com/publication/education-learning-pubs/literacy-and-language-arts-pubs/research-brief-impact-of-the-national-writing-projects-college-ready-writers-program-on-teachers-and-students/

Graham, S., & Perin, D. (2007). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools - A report to Carnegie Corporation of New York.* Alliance for Excellent Education. https://www.carnegie.org/publications/writing-next-effective-strategies-to-improve-writing-of-adolescents-in-middle-and-high-schools/

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education, 43,* 277–303. https://doi.org/10.3102/0091732X18821125

Liu, S., & Kunnan, A. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *Calico Journal, 33*(1), 71–91. https://www.jstor.org/stable/calicojournal.33.1.71

National Center for Educational Statistics (2012). *The nation's report card: Writing 2011* (NCES 2012-470). Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20,* 53–76. http://dx.doi.org/10.1016/j.asw.2013.04.001

Young, V. M., & Kim, D. H. (2010). Using assessments for instructional improvement: A literature review. *Education Policy Analysis Archives, 18*(19). Retrieved Nov 15, 2022, from http://epaa.asu.edu/ojs/article/view/809

Appendix Table 1. Prompt Topics

Topic	Prompt language given to students
Youth activism	How should young people be advocating for change? Youth activism has been growing for many years. Around the world, young people are speaking up and demanding change on issues from climate change to gun violence to social inequality. These activists use a variety of methods to make their voices heard. What is the most effective method for creating change? After reading the provided articles, write an essay that argues the best method young people should use to advocate for change. Defend your claim using clear reasons and relevant evidence from the sources provided, and be sure to acknowledge and address counterclaims to your position.
Criminal Justice	How should we change the court system to make it equitable for everyone? The US criminal justice system, from the courts to policing, does not serve everyone equally. Politicians and activists believe in different ways to improve it. Many believe the problem starts in the courtroom. Mandatory minimum sentencing and high bail amounts have led to a growing prison population, especially for people of color. What is the best way to seek out reform? Your senator is considering introducing legislation to reform the court system. Write an argumentative essay recommending what should be done to make the courtroom fair and equal for everyone. Use evidence from the sources to defend your claim, and be sure to acknowledge and address counterclaims to your position.
Rising Sea Levels	Who should be making decisions on how we protect communities facing unequal impacts of rising sea levels? The effects of a changing climate are not felt equally by all people. Often, communities of color and low-income communities suffer the most from pollution and environmental changes. This is true for communities along the US coast that are facing excessive impacts of flooding caused by rising sea levels and warming temperatures. What is the most effective way to protect these communities? Who should make decisions about how we address sea level rise? Write a letter to the president arguing who you think should be making decisions about protecting communities facing unequal impacts from sea level rise. Use evidence from the sources to defend your claim, and be sure to acknowledge and address counterclaims to your position.
Graffiti	Is graffiti art or vandalism? The city of Covina is preparing to write a position statement on whether graffiti is an art form, or whether it is vandalism. The mayor has invited the public to join in the debate before the city writes its position statement. After reading the provided articles and viewing the video on the topic, write an argumentative, multi-paragraph essay that addresses the question “Is graffiti art or vandalism?” You must support your position with evidence from the texts and video.

Topic	Prompt language given to students
Cell Phones	In the 21st century, information, conveniences, social groups and entertainment are at hand. Quite literally, you can hold in your hand a device that makes all your thoughts and desires come to life through your cell phone. The smart-phone can also be used as a tool to enhance the educational process. This modern convenience is often taken away when students walk into a classroom. Write a letter to the principal that argues whether students should or should not be allowed cell phone use in class. In your research, you have found two videos and two articles on cell phones. Be sure to take notes on each of the sources to gather evidence for your argument, and use those sources to support your position.
Screen Time	As technology has become more common in our daily lives, humans are interacting with computer screens at a higher rate than ever before. Portable devices like laptops, tablets, handheld gaming systems, and especially smartphones have remarkably increased the amount of time teenagers are staring at computer screens. As a result, the American Association of Pediatrics (AAP) recommended a limit of two hours of screen time per day for teenagers. After examining the potential benefits and risks of screen time in the sources provided, write an essay arguing whether or not the AAP should keep the recommended two-hour limit of daily screen time for teenagers or eliminate it. Defend your position using clear reasons and relevant evidence from the sources provided, and be sure to acknowledge and address counterclaims to your position.

Appendix Table 2. Teacher Convening Essay Dataset

Source teacher	Grade	Gender	Race	Free/ reduced lunch %	Total essays	Is It Art?	Cell Phones	Screen Time	Young Activists
1	8	F	White	51–75%	8	4	4		
2	8	F	White	51–75%	8	4	4		
3	8	F	White	51–75%	8			4	4
4	7	F	White	51–75%	9	4	5		
5	8	F	White	>76%	8	8			
6	8	F	White	>76%	8		4		4
7	8	F	White	51–75%	8			6	2
8	7	F	Black	>76%	9		1	8	
9	8	F	White	51–75%	7			3	4
10	7	F	Asian	>76%	9	4	5		
11	8	F	Multi	51–75%	7	7			
12	8	F	White	>76%	8	8			
13	8	F	White	>76%	8	5			3
14	7	F	White	51–75%	9		4	5	
15	8	M	Nat.Am.	51–75%	8	6	2		
16	8	F	White	>76%	5				5
17	8	F	White	>76%	8	7	1		
18	8	F	White	>76%	9		4		5
19	8	F	White	51-75%	9	7	2		
20	8	F	White	>76%	8			8	
21	7	F	White	>76%	9				9

Appendix Table 3. Average Scores by Prompt Topic and Grade (*n* = 3, 233 students)

	Support & Evidence (Baseline | Final)	Claim & Focus (Baseline | Final)	Organization (Baseline | Final)	Language & Style (Baseline | Final)	Overall (Baseline | Final)
By grade level					
Grade 7 (*n* = 1,594)	1.92 | 2.49	1.95 | 2.55	2.09 | 2.72	2.11 | 2.73	8.06 | 10.49
Grade 8 (*n* = 1,641)	2.16 | 2.88	2.24 | 2.97	2.43 | 3.13	2.43 | 3.14	9.24 | 12.12
By prompt topic					
Criminal Justice (*n* = 179)	2.36 | 2.96	2.53 | 3.14	2.65 | 3.22	2.62 | 3.19	10.18 | 12.51
Is It Art? (*n* = 730)	1.74 | 2.49	1.79 | 2.62	1.87 | 2.75	1.97 | 2.77	7.37 | 10.63
Cell Phones (*n* = 1,313)	2.03 | 2.37	2.01 | 2.37	2.18 | 2.53	2.18 | 2.52	8.41 | 9.49
Climate Change (*n* = 53)	2.88 | 3.17	2.92 | 3.52	3.04 | 3.52	3.16 | 3.70	12.00 | 13.91
Screen Time (*n* = 873)	2.08 | 2.60	2.18 | 2.64	2.36 | 2.84	2.33 | 2.84	8.95 | 10.92
Young Activists (*n* = 1,248)	2.11 | 2.69	2.21 | 2.81	2.45 | 3.02	2.42 | 3.01	9.21 | 11.53

Note: n=3,233 students.

Appendix Table 4. Growth on Key Outcome Dimension-Level Score as Predicted by Corresponding Baseline Dimension-Level Score

	Support & Evidence	Claim & Focus	Organization	Language & Style
Outcome 1: Growth from First Baseline to First Final (*n* = 958)				
1 (vs. 4)	1.04 (0.10)***	1.12 (0.10)***	1.04 (0.08)***	1.12 (0.08)***
2 (vs. 4)	0.69 (0.10)***	0.77 (0.10)***	0.76 (0.08)***	0.75 (0.07)***
3 (vs. 4)	0.33 (0.10)**	0.36 (0.10)***	0.35 (0.07)***	0.38 (0.07)***
Outcome 2: Growth from Second Baseline to Second Final (*n* = 434)				
1 (vs. 4)	1.28 (0.19)***	1.29 (0.17)***	1.20 (0.14)***	1.34 (0.13)***
2 (vs. 4)	0.80 (0.18)***	0.81 (0.16)***	0.77 (0.12)***	0.83 (0.11)***
3 (vs. 4)	0.39 (0.18)*	0.35 (0.16)*	0.36 (0.12)**	0.46 (0.11)***
Outcome 3: Transfer from First Baseline to Second Baseline (*n* = 590)				
1 (vs. 4)	1.70 (0.13)***	1.73 (0.12)***	1.88 (0.12)***	1.57 (0.12)***
2 (vs. 4)	1.07 (0.12)***	1.19 (0.11)***	1.40 (0.11)***	1.02 (0.12)***
3 (vs. 4)	0.58 (0.12)***	0.60 (0.11)***	0.77 (0.11)***	0.48 (0.11)***
Outcome 4: Growth from First Baseline to Second Final (*n* = 629)				
1 (vs. 4)	1.77 (0.14)***	1.68 (0.14)***	1.67 (0.13)***	1.57 (0.12)***
2 (vs. 4)	1.28 (0.13)***	1.27 (0.13)***	1.20 (0.12)***	1.07 (0.11)***
3 (vs. 4)	0.75 (0.13)***	0.64 (0.13)***	0.66 (0.11)***	0.50 (0.11)***