Can AI Support Writing Skill Development?

Pilot Evidence from TrueMark's AI Enhanced Writing Platform

by Christina Steiner, PhD and Claire McCloud, EdM

June 2026

Funded by IES Small Business Innovation Research (SBIR) Program, Phase IB,
Award Number: 91990025C0079

Table of Contents

Executive Summary

This report details the findings from an exploratory evaluation of TrueMark's AI-powered writing platform funded through the Institute of Education Sciences (IES) Small Business Innovation Research (SBIR) Phase IB program. TrueMark is an AI-enhanced writing platform targeting middle school teachers and students. The platform enables teachers to create custom writing assignments with built-in Socratic AI tutoring aligned to their rubrics and course materials. Students complete their work in a secure, Google Docs-like environment. Key platform features include a Socratic AI Copilot designed to support student learning, and a Rubric Checker, which compares student work against the teacher's rubric to provide tailored feedback to improve student grades.

The study was conducted across diverse geographies and contexts. It incorporates findings from teacher interviews, student surveys, and platform usage data. The pilot began in February 2026 and concluded in May 2026. The primary innovation that was tested was a student-facing personalized AI Copilot. Additionally, teachers were asked to reflect on a classroom data dashboard to be built out in Phase II, which relies on TrueMark's auto-grading system.

Overall usage of the platform among pilot teachers and students was high. During the pilot, participating students completed an average of 16.2 assignments and initiated an average of 2.7 rubric checks and 2.2 AI Copilot chats per submission. Teachers initiated the assignments and were asked to allow students to use the AI Copilot. However, teachers retained the ability to use their professional judgment to set limits on rubric checks and AI Copilot interactions throughout the study. The median student completed 20 submissions during the pilot or 2-3 assignments/week, demonstrating sustained engagement.

Usability & Feasibility

All eight teachers who were interviewed found core features intuitive and easy to integrate. The Socratic AI Copilot was praised for democratizing student support, the AI grading comments reduced teacher workload, and the platform's detection of student AI usage was the most universally cited driver of adoption. The personalized Socratic AI Copilot's design was described as both pedagogically appropriate and a model for responsible AI use. Six of eight teachers identified onboarding gaps around advanced features. They requested subject-specific training, demo environments, and asynchronous video resources.

Areas for refinement emerged:

1) Rubric auto-grading accuracy: too generous at higher grade levels, too strict at lower levels, with phantom errors undermining teacher trust.

2) AI Copilot grade-level differentiation: advanced students need an "extra low" helpfulness mode to preserve productive struggle.

3) In-app examples of the skills dashboard prompted teachers to request drill-down capability, and transparency into how proficiency levels are calculated.

Early signals of effectiveness also emerged. Students agreed that TrueMark improved their writing across all five competency areas (post-only mean = 3.87/5, $p < .001$). Skill estimates improved significantly across all five dimensions (all $p < .001$), and proficiency level shifts favored upward movement, with downward movement never exceeding 2.5%. A clear dosage-response pattern emerged: more assignment updates predicted more growth, even after controlling for starting skill level.

Seven of eight teachers reported concrete observations of student progress. For example, one teacher reported that students demonstrated enhanced organization in their essay writing on paper and pencil assignments post-TrueMark, which was attributed to the platform's AI Copilot tutoring. Overall, the results were promising, and the analysis uncovered small modifications to be addressed in Phase II prior to formal outcomes testing.

Introduction

TrueMark is an AI-enhanced writing platform for education that enables teachers to create custom writing assignments with built-in AI tutoring aligned to their rubrics and course materials. Students complete their work in a secure, Google Docs-like environment where they can receive instant feedback, guided by teacher-defined criteria. For students, key platform features include a Socratic AI Copilot designed to support student learning by asking questions, never answering. The platform also features a Rubric Checker, which compares student work against the teacher's rubric to provide tailored feedback to improve student grades. From the teacher perspective, the platform includes comprehensive monitoring tools to track student progress, review internal and external AI interactions, detect academic dishonesty, and grade assignments with AI-assisted suggestions. All AI interactions can be customized to support teacher-specified learning objectives while maintaining academic integrity. TrueMark launched commercially in September 2024 and within its first four months was in use across five schools with 1,600 end-user accounts.

In the Spring of 2026, TrueMark partnered with Quantily, LLC to conduct a pilot study of the early effectiveness of platform features. This report organizes key themes identified in the pilot study through the following data sources: (1) Pre- and Post Surveys of students, conducted in mid-February 2026 and May 2026; (2) In-App Reflections of students, conducted multiple times throughout the study; (3) Estimates of students' proficiency levels, conducted at the beginning and end of the study; and (4) Interviews of pilot teachers, conducted between April-May 2026. The report concludes with recommendations for the next Phase of TrueMark's development.

TrueMark's SBIR Phase IB Award

The Small Business Innovation Research (SBIR) program, administered by the Institute of Education Sciences (IES), funds the development and early-stage testing of promising education technology products in two phases: Phase IA supports initial prototype development, and Phase IB supports continued product refinement and a small-scale pilot study to evaluate usability, feasibility, and early effectiveness. TrueMark was awarded an IES SBIR Phase IB grant to build on its existing platform and study its impact in real classrooms. Phase IB enhancements included a personalized feedback loop that adapts AI interactions to individual teacher grading patterns and student needs, and the backbone for a skills dashboard that tracks student proficiency across five writing dimensions, e.g., Organization and Structure, continuously over time. The purpose of the pilot study was to assess how well teachers and students could navigate and benefit from the platform's new features, and generate independent evidence to inform a Phase II proposal for a more rigorous, larger-scale study.

Research Questions

Students

1) **RQ1, Usability:** How intuitive and user-friendly do students find the platform and its personalized AI co-pilot function?
2) **RQ2, Feasibility:** Do students receive adequate training to use the platform effectively, and what additional support might enhance their engagement?
3) **RQ3:** To what extent do students perceive the product as helpful in improving their writing skills and overall learning experience?

Teachers

4) **RQ4, Usability:** How intuitive and easy is it for teachers to navigate and implement the product's features, including student-specific features? Does usability vary by teacher experience?
5) **RQ5, Feasibility:** To what extent do teachers have the necessary resources, training, and time to effectively integrate the product into their instruction?
6) **RQ6, Barriers:** What barriers do teachers encounter in using the product, and how do these impact their ability to improve instructional strategies for writing?
7) **RQ7, Effectiveness:** To what extent do teachers perceive the product as helpful in improving students' writing skills and overall learning experience?

Methods

Data were analyzed for both teacher and student components of the study using a mixed-methods approach. Qualitative data, including teacher interview transcripts and open-ended survey responses, were analyzed using applied thematic analysis to identify salient patterns and themes (Guest, MacQueen, & Namey, 2012). Multiple members of the research team coded qualitative data, and interrater reliability ensured consistency in theme identification. Quantitative analyses included descriptive statistics for survey responses, administrative platform data, and instructional usage metrics. Analyses focused on summarizing patterns of use, frequencies, and distributions across teacher and student data. Where appropriate, correlational analyses were conducted to examine associations between student usage patterns and teacher instructional data. All analyses were descriptive and correlational in nature.

Data Sources

To understand the answers to the research questions, researchers collected and analyzed: (1) student data from 50 matched pre/post surveys, (2) data from pilot and comparison student users who completed brief, in-app surveys multiple times throughout the pilot, (3) student prototype

usage and skills data, and (4) data from eight interviews with teacher users. Table 1 describes each data source, the relevant research questions, timing of data collection, sample size, and analysis type.

Table 1. Research Questions, Sample Sizes, and Data Sources, Collection, and Analysis

Target and Research Question	Data Type	Timing	n	Analysis Type
Students; RQ1, RQ2, RQ3	Pre/Post Survey	Mid Feb 2026–End of March 2026	50 Matched	Quantitative pre/post t-tests; Qualitative thematic analysis
Students; RQ1, RQ3	In-App Reflections	Ongoing, after every third assignment	693-675 overall; ~25 with repeated ranking to computer change scores	Quantitative; Treatment vs. Control
Students; RQ3	Skill Estimates and Usage Data	Pilot start vs. End	~300 with pre and post skills estimates	Quantitative pre/post t-tests; OLS regression
Teachers; RQ4-RQ7	1:1 Interviews	April-May 2026	8	Qualitative thematic analysis

Consented Analytic Sample

The analysis leverages secondary platform data as well as survey data from a subset of students in pilot schools.

Table 2. Pilot Schools, Consent Counts, Classes Taught

School	Type, State	Classes Taught	Number of Consented Teachers	Number of Consented Students Who Completed Pre/Post Surveys
Pinecrest Cadence	Charter, Nevada	English Language Arts	2 Teachers	12 Students
Scotch Plains-Fanwood	Public, New Jersey	Language Arts	2 Teachers	15 Students
Ursuline	All Girls Private, New York	English & Professional Development	3 Teachers	4 Students

School	Type, State	Classes Taught	Number of Consented Teachers	Number of Consented Students Who Completed Pre/Post Surveys
Xavier	All Boys Private, New York	Global Studies & Religion 9	2 Teachers	19 Students
Total			**9 Teachers**	**50 Students**

Note. One teacher was unable to schedule an interview due to circumstances outside of the control of the study.

Analysis of Student Data

Overall usage of the platform among pilot teachers and students was high. During the pilot, participating students completed an average of 16.2 assignments and initiated an average of 2.7 rubric checks and 2.2 AI Copilot chats per submission. Teachers initiated the assignments and were asked to allow students to use the AI Copilot. However, teachers retained the ability to use their professional judgment to set limits on rubric checks and AI Copilot interactions throughout the study. The median student completed 20 submissions during the pilot or 2-3 assignments/week, demonstrating sustained engagement.

Student Survey

To assess whether student perceptions changed over the course of the pilot, a pre/post survey was administered to a subset of students in pilot classrooms. Participation was voluntary and required active parental consent. Ultimately, 50 students and parents consented and participated in the research survey, which was sufficient for the aims of this small pilot project.

The pilot occurred in the context of a private-sector edtech platform operating across regions and school-types (charter and private), and involved navigating institutional approval structures, school communication protocols, and parent outreach timelines that were inherently more complex than those typical of a district-run study. This is consistent with the broader implementation literature on consent in school-based research, where response rates are frequently affected by the layered communication channels between program operators, researchers, school administrators, and families. Overall, the consent and communication infrastructure established during this pilot will inform a more streamlined process in Phase II, with the expectation of substantially higher participation rates.

Six constructs were measured pre and post. (1) Usability and (2) Feasibility were assessed using nine items adapted from the System Usability Scale to reflect the specific features and classroom context of the TrueMark platform (Brooke, 2013). (3) Writing self-efficacy was assessed using six items measuring core writing competencies including ideation, organization, evidence use, conventions, revision, and persistence, with the conceptual structure and response format aligned

to the Self-Efficacy for Writing Scale (Bruning et al., 2013). (4) Perceived writing improvement was measured using five researcher-developed items targeting core writing competencies aligned with the writing self-efficacy construct, allowing for conceptual triangulation across the two scales. Finally, to understand TrueMark's impact on student perceptions of AI as a learning tool, Human-Centric Learning and (6) Confidence with AI were measured directly using two subscales from the recently developed Student AI Competency Self-Efficacy Framework (Chiu et al., 2025).

One construct, AI Copilot Perceptions, was measured in the post-survey only because students needed to have used the AI Copilot in order to meaningfully respond to the items. The five items were researcher-developed and grounded in the theoretical and design foundations of the TrueMark platform. Specifically, items were developed to assess student perceptions of feedback actionability and prioritization (Kehrer, Kelly, & Heffernan, 2013), feedback personalization and tailoring (Kabudi, Pappas, & Olsen, 2021), confidence in revision, and the platform's Socratic, autonomy-preserving approach to feedback (Delić & Bećirović, 2016; Ho, Chen, & Li, 2023; Fabio, Plebe, & Suriano, 2024).

Instrument Reliability

Reliability analyses confirmed that all six constructs demonstrated acceptable to good internal consistency in this sample. As shown in Table 3, Cronbach's alpha values ranged from 0.79 to 0.88, with all constructs meeting or exceeding the acceptable threshold ($\alpha \geq .70$). These reliability estimates provide independent psychometric evidence that the survey instruments functioned as intended within the consented pilot sample.

Table 3. Internal Consistency Reliability (Cronbach's α) for Each Construct

Construct	# of Items	Cronbach's α	Interpretation
Usability	5	0.80	Good
Feasibility	4	0.81	Good
Writing Self-Efficacy	6	0.79	Acceptable
Perceived Writing Improvement	5	0.85	Good
AI Copilot Perceptions	5	0.88	Good
Human-Centric Learning	3	0.80	Good
Confidence with AI	4	0.86	Good

Note. Reliability values ≥ .70 are considered acceptable, ≥ .80 good, and ≥ .90 excellent.

Perceived Writing Improvement (Post-Only)

Because the Perceived Writing Improvement construct asked students to reflect on TrueMark's impact on their writing, this scale was administered at post-test only and could not be analyzed as a change score. Among the 50 consented students who completed the post-survey, mean agreement on this construct was 3.87 on a 1–5 scale, indicating that on average, students agreed that TrueMark

helped them improve the overall quality of their writing, write more clearly, organize their ideas better, add stronger details or evidence, and revise more effectively. For example, students answered the question, "The feedback I received from TrueMark feels tailored to my specific writing needs." A one-sample t-test against the neutral midpoint (3 = "Neither agree nor disagree") confirmed this finding was statistically significant ($p < .001$). Students affirmatively agreed that TrueMark improved their writing across all five competency areas.

Construct-Level Pre-Post Change Scores

At the construct level, five of six constructs showed positive mean change scores, indicating that student perceptions trended in a favorable direction across Usability (average change = +0.118), Feasibility (average change = +0.060), Writing Self-Efficacy (average change = +0.019), AI Copilot Perceptions (average change = +0.108), and AI Confidence (average change = +0.095). Human-Centric Learning showed slightly negative mean changes (average change = -0.014). None of the construct-level change scores reached statistical significance, which is consistent with the relatively small, matched sample (n = 49–50) and the limited statistical power to detect small effects. The consistent positive direction across five of six constructs is encouraging and suggests that with a larger sample, meaningful effects may emerge.

Table 4. Construct-Level Pre-Post Change Scores

Construct	n	Mean Change	SD
Usability	50	+0.118	0.565
Feasibility	50	+0.060	0.533
Writing Self-Efficacy	50	+0.019	0.636
AI Copilot Perceptions	50	+0.108	0.681
Human-Centric Learning	49	-0.014	0.742
AI Confidence	50	+0.095	0.737

Note. Mean change scores reflect post minus pre construct means. Constructs measured on 1–5 scales. All p-values reflect two-tailed tests against zero.

Item-Level Findings

At the item level, two items reached statistical significance. Students showed a significant positive change in their agreement that TrueMark has features that help them complete writing assignments ($p = .001$, average change = +0.38). Students also reported a significant increase in agreement that the feedback they receive from TrueMark helps them prioritize the most important changes to improve their writing ($p = .044$, average change = +0.24). This latter finding is particularly meaningful as it maps directly to a core design intention of the platform: providing personalized, actionable feedback that guides student revision.

Table 5. Statistically Significant Item-Level Pre-Post Change Scores

Construct	Item	n	Mean Change	SD	t	p
Usability	TrueMark has features that help me complete my writing assignments.	50	+0.38	0.725	3.705	.001
AI Copilot Perceptions	The feedback I receive from TrueMark helps me prioritize the most important changes to improve my writing.	50	+0.24	0.822	2.064	.044

Note. Mean change scores reflect post minus pre item means. Items measured on 1–5 scales. p-values reflect two-tailed tests against zero.

Two additional items showed trends approaching significance. Students tended toward greater agreement that TrueMark is easy to use (p = .058, average change = +0.20) and that they find it easy to learn AI technology (p = .079, average change = +0.26). Taken together, these findings suggest that students are registering TrueMark's core value proposition, i.e., actionable, prioritized feedback, and are becoming more comfortable with the platform and with AI-based learning tools more broadly. This matches student comments that demonstrate the range of reaction to TrueMark's intentional Socratic design, with some students expressing frustration at not receiving direct answers and others embracing the guided approach:

> *"Make it give more information, and for it to stop asking me the questions because after all it exists for humans to question it, and not for it to question us."*
>
> *"Copilot sucks it didn't write my essay."*
>
> *"I liked how it was helping me, not giving me answers, and gave me feedback on what I need to fix."*

The pattern of findings across both significant and trending items points to a consistent signal: students who used TrueMark showed meaningful growth in perceiving its feedback as actionable, targeted, and conducive to independent thinking.

Open-Ended Survey Responses

Four open-ended post-survey questions were coded thematically by two independent coders. Coding was conducted in two rounds: an initial independent round (Round 1) followed by a consensus discussion to reconcile discrepancies (Round 2). The unit of coding was the individual student response (n = 50). Inter-rater reliability (IRR) was calculated as percent agreement using the formula: 1 − (|C − Cl| / average (C, Cl)) for each theme. Final counts (IRR=97%) from Round 2 are used in Table 6.

Table 6. Final Thematic Counts and Interpretations by Question

Theme	Final Count (n=50)	Interpretation
Q35: Best feedback received from AI Copilot?		
Evidence	12 (24%)	The most common feedback students cited was guidance on strengthening evidence and connecting it to their claims, which directly aligned with the skill area showing the strongest quantitative growth.
Clarity	11 (22%)	Students frequently described receiving feedback on making their writing clearer and easier to understand, including sentence-level revisions.
Sentence Structure/Grammar	13 (26%)	A smaller subset described specific feedback on sentence construction and grammar.
Organization/Flow	8 (16%)	A smaller subset described feedback on transitions and paragraph-level organization.
Q36: To what extent has TrueMark changed how you write?		
Proofreading/Catching Mistakes	18 (36%)	The most common response described TrueMark as improving how students review and catch errors in their own work; a revision behavior change rather than a writing behavior change.
No Change	15 (30%)	Students reported that TrueMark had not changed how they write, though several qualified this by noting it helped them in the moment or with specific assignments.
Structure/Organization	12 (24%)	Students described TrueMark as influencing how they organize their writing and structure their paragraphs.
Q37: Favorite feature in TrueMark?		
Rubric Checker	31 (62%)	The rubric checker was overwhelmingly the most frequently cited favorite feature, with students describing it as a tool for understanding expectations and monitoring their own progress in real time.
AI Copilot	14 (28%)	The AI Copilot was the second most cited favorite feature, with students describing it as useful for targeted questions and writing-specific guidance.
Q38: One feature you wish you could change?		
No Changes Desired	13 (26%)	A subset of students reported satisfaction with the platform as-is and did not identify anything they would change.
Frustration with Socratic Method	14 (28%)	The most common change request was for the AI Copilot to provide more direct answers rather than asking guiding questions, an intentional design element.
Desire for More Detailed Feedback	12 (24%)	Students expressed a desire for deeper, more specific explanations rather than general guidance, another intentional design element.

Theme	Final Count (n=50)	Interpretation
Reliable Spell Check	5 (10%)	Students noted that the spell check feature was inconsistent or unreliable, flagging this as a usability concern.

Note. Final counts reflect Round 2 consensus coding. Percentages reflect proportion of total respondents (n = 50). Responses could be coded for multiple themes; totals within a question may exceed n.

Across all four open-ended items, four findings stand out:

1) The incorporation of evidence-based writing was the most cited area of AI Copilot feedback (24%), consistent with the quantitative finding that skill growth was strongest in Analysis & Reasoning and Evidence & Support (detailed below).

2) The rubric checker dominated as students' favorite feature (62%), a preference that was stable across both pre and post surveys and was already students' top choice before the pilot began. The AI Copilot was also mentioned frequently as a favorite feature (28%).

3) Frustration with the AI Copilot's Socratic approach was the most common change request (28%), with students wanting direct answers rather than guiding questions. As noted in the quantitative findings, this friction is consistent with the platform functioning as intended and should be interpreted as evidence of engagement with the feedback model rather than a product failure.

4) A meaningful proportion of students (30%) reported that TrueMark had not changed how they write, though most of these responses described changes to their revision process rather than initial writing. This distinction between writing and revising is important and it may be worth testing more directly in Phase II.

The most meaningful finding from the open-ended pre/post comparison is that the nature of how students described their experiences became more specific and behaviorally concrete over the course of the pilot. At pre, student responses tended to be general and procedural, describing TrueMark in broad terms: "it helps me edit my writing," "it helped me notice mistakes easier," "it gives me feedback." At post, students were more likely to name specific skill areas, describe feedback interactions, and articulate what they did with the feedback they received. For example, at pre, a student described receiving feedback as: "It gave me feedback on how to improve my writing and how to make people better understand my writing." By post, responses were more targeted: "The best feedback I received was what I could think about to help connect my quote back to my claim and better explain it. With this I went back to my work and knew what to add on and look for when doing so." This shift from general descriptions of feedback to specific writing strategies suggests that students developed a more sophisticated understanding of how to use TrueMark's feedback over time.

Students attributing grade improvement to TrueMark tripled from pre (5%) to post (16%). While self-reported grade outcomes were not independently verified, the increase in frequency and specificity of these accounts over the pilot is notable. In the post survey, students described acting on specific feedback, strengthening evidence, and clarifying claims. Selected quotes illustrating this theme include:

> *"So Truemark told me to write with more details so I did and I got the best grade on my essay that I ever had."*
>
> *"Copilot would tell me how to explain my sentences more clearly and better. I used it to boost my grade."*

Student Perceptions of the AI Copilot in Pilot vs. Non-Pilot Classrooms (In-App Data)

Students completed in-app survey items more than once during the pilot, enabling a preliminary look at whether perceptions of the AI Copilot changed over time and in pilot versus non-pilot classrooms. Change scores were calculated for two items: 1) perceived helpfulness of the AI Copilot and 2) perceived impact of the AI Copilot on writing. The specific questions were, "How helpful was the AI Copilot feedback?" and "How much did the AI Copilot feedback help you improve your draft?" Response categories were "Not at all helpful", "A little helpful", "Helpful", and "Very helpful". The change in scores during the pilot were coded as -1 (declined), 0 (no change), or +1 (improved). Among students who did not experience TrueMark's new features, average change scores were slightly negative for both helpfulness (average change = -0.11) and writing improvement (average change = -0.16), suggesting modest decline over time. Treatment students showed the opposite pattern, with small positive change scores on both helpfulness (average change = +0.12) and writing improvement (average change = +0.17). Independent samples t-tests found these differences to be in the expected direction but not statistically significant (helpfulness: $p \approx .23$; writing improvement: $p \approx .11$).

Similar to the feedback requests users encounter after using most software products, the in-app survey items were delivered as optional pop-up prompts, and most students did not respond. A low response rate was expected with this type of voluntary, unannounced prompt. As a result, the subsample available for change-score analysis was small, limiting statistical power. However, the consistent directional pattern across both items with pilot students reporting improved perceptions and non-pilot students declining is nonetheless noteworthy and warrants further investigation with a larger longitudinal sample.

Student Skill Growth During the TrueMark Pilot

To assess whether students demonstrated growth in writing skills over the course of the pilot, skill estimate scores were analyzed at the individual student level from the beginning to the end of the pilot period. Each student was assessed on up to five writing skills, with skill categories determined

by their classroom teacher and broadly corresponding to five thematic areas: Claim & Thesis Development, Evidence & Support, Analysis & Reasoning, Organization, and Language & Style. Students with zero recorded assignment updates for a given skill were excluded from the analysis. Two complementary outcomes were examined: continuous skill estimate scores (ranging from 0 to 1) and categorical proficiency levels (Developing, Proficient, Advanced).

Continuous Skill Estimate Growth

One-sample t-tests were used to determine whether mean growth in skill estimates differed significantly from zero from the beginning of the pilot to the end of the pilot. Among students with multiple completed assignments that were used to compute skill-scores, statistically significant growth was observed across all five skills areas ($p<.001$, Table 7). Mean growth scores ranged from +0.018 to +0.024, with sample sizes ranging from 369 to 417. The consistency of this finding across every skill dimension represents a compelling quantitative finding of the pilot. Every skill area improved, and the probability that these findings occurred by chance is extremely low. Skill 3 (Analysis & Reasoning) showed the strongest growth signal, followed closely by Skill 2 (Evidence & Support).

Table 7. Continuous Skill Estimate Growth by Skill

Skill	Thematic Area	N	Average Growth	t stat	95% CI	P value
Skill 1	Claim & Thesis	414	+0.023	11.13	[.019, .027]	<.001
Skill 2	Evidence & Support	369	+0.024	13.27	[.021, .028]	<.001
Skill 3	Analysis & Reasoning	411	+0.022	14.19	[.019, .025]	<.001
Skill 4	Language & Style	417	+0.017	10.62	[.015, .021]	<.001
Skill 5	Organization & Structure	407	+0.019	10.88	[.015, .022]	<.001

Note. Average growth reflects change in skill estimate score (scale 0–1) from pilot start to end. All p-values reflect two-tailed tests.

Proficiency Level Change

Proficiency level changes were examined as a complementary outcome, coded as moved down (−1), no change (0), or moved up (+1). One-sample t-tests against zero were used to assess whether the mean direction of change was significantly positive. Four of five skills showed statistically significant upward movement in proficiency level (Table 8). Most students maintained their proficiency level across the pilot period, which is expected given that most students entered the pilot already rated as Proficient. Significant upward shifts were observed for Claim & Thesis ($p < .001$), Evidence & Support ($p = .028$), Analysis & Reasoning ($p < .001$), and Language & Style ($p = .049$), Organization & Structure did not reach significance ($p = .194$), though the directional pattern remained positive.

Upward proficiency changes consistently and substantially outpaced downward changes across all five skills. Downward movement was rare across all skill areas, never exceeding 2.5% of students.

Table 8. Proficiency Level Change by Skill Slot

Skill	Thematic Area	N	Moved Down	No Change	Moved Up	T stat	P value
Skill 1	Claim & Thesis	414	5 (1.2%)	375 (90.6%)	34 (8.2%)	4.76	<.001
Skill 2	Evidence & Support	369	7 (1.9%)	344 (93.2%)	18 (4.9%)	2.21	.028
Skill 3	Analysis & Reasoning	411	5 (1.2%)	368 (89.5%)	38 (9.3%)	5.19	<.001
Skill 4	Language & Style	417	6 (1.4%)	396 (95.0%)	15 (3.6%)	1.97	.049
Skill 5	Organization & Structure	407	11 (2.7%)	378 (92.9%)	18 (4.4%)	1.30	.194

Note. Proficiency levels coded as Developing (1), Proficient (2), Advanced (3). Change scores reflect end minus start proficiency level.

Number of Updates and Skill Growth

To examine whether greater platform engagement was associated with greater skill growth, Pearson correlations and simple OLS regressions were computed between each skill's number of assignment updates and its growth score. All five associations were positive and statistically significant (all $p < .001$). Correlations ranged from $r = 0.26$ (Organization & Structure) to $r = 0.53$ (Evidence & Support), a moderate dosage-response relationship (Table 9). Regression coefficients indicate that each additional assignment update was associated with approximately +0.004 to +0.009 points of growth on the skill estimate scale.

Evidence & Support showed the strongest association ($r = 0.53$). This finding is consistent with the broader literature on deliberate practice and formative feedback and provides preliminary correlational support for the platform's potential causal impact on writing skill development (Ericsson, Krampe, &Tesch-Romer, 1993).

It should be noted that the number of updates across skill areas was highly intercorrelated ($r = 0.82–0.99$) because most assignments contained all five skills. The relationship between the number of skills updates per student partially reflects volume of assignments rather than skill-specific practice. Phase II should include multivariate models that account for classroom fixed effects and an external baseline like a nationally normed benchmark to precisely isolate the contribution of platform engagement to skill growth.

Table 9. Association Between Assignment Updates and Skill Growth

Skill	Thematic Area	r	R^2	Coeff.	Interpretation	P value
Skill 1	Claim & Thesis	0.47	0.23	+0.009	Moderate	<.001
Skill 2	Evidence & Support	0.53	0.28	+0.007	Moderate-Strong	<.001
Skill 3	Analysis & Reasoning	0.28	0.08	+0.005	Moderate	<.001
Skill 4	Language & Style	0.36	0.13	+0.005	Moderate	<.001
Skill 5	Organization & Structure	0.26	0.07	+0.004	Weak-Moderate	<.001

Note. r = Pearson correlation between number of assignment updates and skill growth. R^2 from simple OLS regression. Coeff. = unstandardized regression coefficient (growth per additional update). All p < .001.

Accounting for Baseline Skills

To examine whether greater platform engagement was associated with greater skill growth, OLS regressions were estimated for each skill with number of assignment updates and starting skill estimate as predictors. Controlling for starting skill level is important because students who begin lower have more room to grow, and because lower-starting students may also complete fewer assignments.

The update effect was positive and statistically significant across all five skills after controlling for starting skill level (Table 10). Once baseline skill level is held constant, each additional assignment is associated with even more growth. Regression coefficients indicate that each additional assignment update was associated with approximately +0.005 to +0.012 points of growth on the skill estimate scale, net of starting level.

The starting skill estimate was negative and significant across all five skills, confirming that students who began the pilot at lower skill levels showed greater growth. For example, a simple interpretation of skill 1 is: for each 1-unit increase in starting skill estimate, predicted growth decreases by 0.213 points. This suggests that TrueMark is providing meaningful guidance to students who enter writing at a low level. Phase II should extend this model by increasing the sample size, including a comparison group, adding classroom fixed effects, and, if available, incorporating student-level covariates to further isolate the platform's contribution to skill growth. Additionally, the impact of TrueMark should be tested with an outcome measure that is exogenous to the system, e.g., a standardized writing assessment.

Table 10. Multivariate OLS Regression: Assignment Updates and Starting Skill Predicting Growth

Skill	Thematic Area	N	Updates Coeff.	Start Estimate Coeff.	p
Skill 1	Claim & Thesis	414	+0.012***	-0.213***	<.001

Skill	Thematic Area	N	Updates Coeff.	Start Estimate Coeff.	p
Skill 2	Evidence & Support	369	+0.010***	-0.101*	<.001
Skill 3	Analysis & Reasoning	411	+0.007***	-0.101*	<.001
Skill 4	Language & Style	417	+0.007***	-0.114***	<.001
Skill 5	Organization & Structure	407	+0.005***	-0.090***	<.001

*Note. Updates Coeff. = unstandardized OLS coefficient for number of assignment updates. Start Estimate Coeff. = unstandardized coefficient for starting skill estimate (scale 0–1). *** p < .001.*

Key Insights from Analysis of Student Findings

1) Five of six survey constructs showed positive mean change scores pre-to-post (Usability +0.12, Feasibility +0.06, Writing Self-Efficacy +0.02, AI Copilot Perceptions +0.11, AI Confidence +0.10; Human-Centric Learning -0.01), with none reaching statistical significance in the matched sample (n = 49–50). At the item level, two items reached significance: students showed increased agreement that TrueMark has features that help them complete writing assignments (p = .001, +0.38) and that TrueMark feedback helps them prioritize the most important changes to their writing (p = .044, +0.24), the latter mapping directly to the platform's core design intention of actionable, personalized feedback. Students affirmatively agreed with the perception that TrueMark improved their writing across all five competency areas assessed, with a post-only mean of 3.85/5.

2) Writing skill estimates improved significantly across all five dimensions tracked by the platform, with growth observed from the beginning to the end of the pilot period (all p < .001) and students moving up in proficiency consistently outnumbering those moving down.

3) A clear dosage-response relationship emerged: students who completed more assignments on TrueMark showed greater skill growth across all five areas, even after controlling for starting skill level, which suggests the platform's benefit compounds with engagement.

4) The rubric checker was students' most-used and most-valued feature (62% named it their favorite), and growth was strongest in the skill areas most directly supported by the rubric checker during the revision process: Analysis & Reasoning and Evidence & Support.

5) Over the course of the pilot, students' descriptions of their TrueMark experience became more specific and behaviorally concrete, shifting from general statements about feedback to naming particular skill areas and describing how they acted on guidance received, a qualitative signal of deepening platform literacy.

6) The most common student change request was for the AI Copilot to provide direct answers rather than asking guiding questions or frustration with the AI Copilot's Socratic design

(52%). This is consistent with the platform functioning as designed and should be interpreted as evidence of engagement with the Socratic feedback model rather than a product flaw.

Analysis of Teacher Data

This section presents key themes from eight TrueMark teacher interviews conducted between April and May 2026. Findings are organized by research question.

Teacher Perceptions of Usability

Overall, all eight teachers who were interviewed found TrueMark's basic features to be intuitive and easy to learn. Less technology savvy teachers were comfortable learning from peers and by trial and error.

Usability Finding 1: The personalized AI Copilot is easy to use.

Across all eight interviews, teachers reported that the student-facing, personalized AI Copilot, which was added as part of the SBIR 1B project, is highly usable. Noting the AI Copilot's foundation in the Socratic method, teachers praised the AI Copilot for encouraging students to ask thoughtful questions without giving them the answers.

> *"I love that it has taken this idea of a helpful AI bot and turned it into a mini teacher that's not just meant to feed you the answers."*

> *"The students have gotten smarter with their questions. They're not asking 'how can you make this better.' They're asking, 'is my intro including all the elements that are on the rubric?' — being very specific and deliberate."*

Teachers also noted that the AI Copilot helps them address student questions more efficiently than without the AI Copilot. For students, the AI Copilot provides targeted, one on one support on demand.

> *"In a classroom of 35 students, I can only help one at a time. Now every single student has access to help at the same time. That's what's been really great."*

> *"The ones who actively engage with it and use it as a tool — they ask me a lot less questions. They themselves appear to be less stressed about writing and more able to complete it in the time allotted. For some of them, that's a big change from where they were first quarter."*

Additionally, multiple teachers noticed timid students, who do not normally ask questions of the teacher, engaging with the AI Copilot. Teachers also appreciated the opportunity to communicate with these more reserved students by leaving encouraging comments or making suggestions within the platform.

> *"I feel like I'm able to reach more students in the amount of time, rather than them coming up to me and just asking those questions. Sometimes they're not always willing to come up and ask questions if they're shy."*
>
> *"Especially the quiet kids…it's like they're finally able to get the questions that they wouldn't necessarily have asked. I like to go into the more quiet ones and see if they're using it, and they're using it. And they ask those questions they wouldn't have asked before."*
>
> *"I don't like to call students out in the classroom… sometimes they feel embarrassed. So I can look at their work on the screen and just leave a little comment—"you're doing great, keep going," or "come up and ask me if you're not sure." I feel like I'm able to reach more students in the amount of time."*

Usability Finding 2: Basic TrueMark Features are intuitive and easy to implement.

Basic features such as assignment creation and AI detection are highly usable and easy to integrate into instruction.

> *"It's pretty easy to create an assignment… if I want to assign something on a whim, I can create it in three minutes and just have students submit it."*
>
> *"Before TrueMark… we were going back to paper-and-pen writing because we just didn't really have a solution. I think this is a good happy medium—they get extra support, even if the copilot and rubric aren't 100% helpful at all times. It's like having a co-teacher."*

Usability Finding 3: AI-Generated grading comments are highly valued for saving teacher time.

The AI-generated comments are highly valuable to all and expedite the grading process. In all eight interviews, teachers reported using TrueMark's AI-generated comments as a foundation for grading, before adding the teacher's own human touch. From there, they adjust and expand upon the feedback, tailored to the specific student's needs and classroom context. For teachers who grade many assignments, the objective AI feedback, aligned with the teacher's rubric, ensures that grading remains fair, consistent, and accurate across students.

> *"It writes for me what I would write, but it's saving me the time of actually having to write it. I love that feature. I think that's fabulous. If that ever went away, I would probably cry."*
>
> *"I like that I have AI grading to be like, center me — is this a B according to how I've been grading it? It helps kind of steer me back into looking at it more objectively."*

> *"I don't have the time to type the same thing 80 times. I'll grade it, but I'll look at the comments, delete any I don't think are relevant, keep any that are, and tell the students: anything you're seeing is my grade, with a mix of TrueMark and my comments."*
>
> *"These are kids who live in a world where everything's immediate. When it's not immediate, like if I had just given them the essay with feedback at the end, I don't think they would pay attention to it. And they're getting it as they're writing, more efficiently than what I could do if I'm trying to do it for 100 papers."*

Teacher Perceptions of Feasibility

Feasibility Finding 1: Personalized AI Copilot's Socratic design makes it feasible for students to engage in critical thinking while using AI.

Overall, teachers reported that they had sufficient resources, training, and time to integrate the AI Copilot into their instruction. Across the board, teachers praised the AI Copilot as a time saver and very easy to use. The AI Copilot's Socratic, boundary-setting design, which encourages students to reach solutions independently rather than directly providing them the answers, was cited as a distinctive feature. When students intentionally probe the AI Copilot to give them the answers or ask off-task questions, teachers characterize the AI Copilot's responses as "sassy," insofar as it disengages from students entirely. Teachers also appreciate this aspect of the AI Copilot for the way that it demonstrates to students what appropriate AI use looks like.

> *"The co-pilot is so sassy. And I absolutely love that."*
>
> *"We don't pretend that AI doesn't exist. We know it exists and we want these kids to know how to use it ethically and responsibly. That, to me, has been the greatest value."*

Feasibility Finding 2: While basic features are easy-to-use, teachers need more guidance on the advanced features.

Usability of TrueMark's advanced features varied greatly by teacher, with 6 out of 8 interviews mentioning the need for additional or more targeted onboarding. For teachers who described themselves as tech-savvy, TrueMark's features felt learnable through experimentation. However, teachers who did not have substantial existing knowledge of AI and technology suggested that a more comprehensive onboarding for advanced TrueMark features is warranted. At the grade or department level, the consequences of under-training could have ripple effects, ultimately causing widespread gaps.

Specifically, teachers requested an onboarding that covers the full suite of TrueMark tools and allows teachers to demo features with an example class and review example data from AI Copilot logs, video-playbacks, typing speed flags, rubric scores, and the skills dashboard.

> *"I think onboarding needs to be very subject-specific, and not only 'here's how you create an assignment step by step' but 'here's how you will receive the assignment*

> *afterward' — from the teacher perspective. And then a walkthrough on how to interpret student interaction with the copilot."*

This implies that TrueMark should add to the current onboarding by offering a more robust training series, a repository of on-demand videos, and webinars to explore advanced and new features to close this gap. Enhanced onboarding and webinars would help teachers discover useful features sooner. Additionally, offering an asynchronous onboarding would ensure teachers who miss group sessions are not left behind about advanced or novel features.

Teacher Reported Barriers

Barrier 1: AI Copilot helpfulness settings should include an "extra low" mode.

The most prominent critique of the AI Copilot, which could be addressed and tested in Phase II, relates to the helpfulness of the AI Copilot's feedback. Currently, teachers can control the level of help (high or low) the AI Copilot provides on each assignment at the classroom level. For many students, the low level of support is too helpful, and leads students to make changes to their writing before they've had a chance to get there themselves.

> *"A student would ask for a synonym, or for another word to use, and then it would answer the question, but then say, "hey, you haven't worked on your conclusion yet." I found that very frustrating because it wasn't allowing them to go through the process of writing and editing… They might get there by themselves. They might not. But give them a chance."*

> *"It's always recommending things. Sometimes it will say great job, and it'll still make recommendations. That's a little confusing for students because they're saying, well, it said I got this on the rubric checker, but it's also telling me I can make it stronger.".*

In response, TrueMark should add an "extra low" AI Copilot helpfulness option, which would be particularly useful for advanced students. Currently, teachers who want to reduce AI Copilot guidance for these high-performing students find the lowest existing setting still tells students specifically what to add or change. With this student-level toggle capability, teachers would ideally be able to adjust the feedback these students receive to be more basic guidance that is slightly mysterious about what is missing, e.g., "you are still missing two key elements." Alternatively, to appropriately scaffold instruction for students who require additional support and a higher level of AI Copilot helpfulness, feedback may look more like "try again by adding these three things," as opposed to the lower-level helpfulness of "try again."

Barrier 2: Rubric Auto-Grading.

Although teachers report that the rubric auto-grading feature is a helpful anchor, all eight teachers interviewed suggested that its accuracy should be improved to increase trust.

Teachers of higher grades often said that the rubric grades student work too generously. For example, AP course teachers reported that TrueMark regularly granted credit for the mere presence of keywords, rather than the quality of students' elaboration. In the lower grades, teachers found the auto-grading to be too strict, often flagging grammatical or structural writing errors that students had not yet been taught. More broadly, teachers noted phantom errors with the auto-grading, where TrueMark flagged issues relating to insufficient word count or lack of MLA formatting that did not exist.

Teachers also indicated that they would benefit from a feature within the auto-grading that learns the teacher's feedback and grading patterns over time. Many students make similar mistakes over time, and teachers would appreciate it if the platform would recognize consistent feedback from teachers and save them time by already mentioning those in its auto-generated feedback. For example, one teacher shared, "I almost wish that it would take my feedback that I give over and over again to these students into consideration. Like, okay, the writing is similar to this and she normally scores that a four. So it would be a four here."

Barrier 3: Accommodations for All Learners.

Several teachers raised the need for accessibility and accommodation features not currently available in the platform: speech-to-text input for students with writing-related learning differences, dark mode for students with sensory sensitivities, and page break controls for students accustomed to typed formatting supports. These barriers did not impede the pilot, but they are meaningful product development considerations for Phase II as TrueMark scales to a more diverse student population.

Teacher Reported Signals of Effectiveness

Effectiveness Signal 1: Teachers noted shifts in instructional strategy.

In 6 out of 8 interviews, teachers noted how TrueMark is enabling concrete instructional changes. One teacher increased student writing volume by 2-3 pieces per marking period. In other interviews, two teachers explained that they grade assignments more harshly when students have access to the AI Copilot and rubric checks, holding students to higher content and structural standards. A fourth teacher reported using TrueMark as a formative study tool to simulate an AP environment, highlighting the lockdown browser. Together, these genuine instructional strategy shifts, enabled by TrueMark platform, signal its impact in real classrooms.

Effectiveness Signal 2: Teachers noted indicators of student progress.

When asked about TrueMark's effectiveness at improving student writing, 7 out of 8 teachers reported meaningful, concrete indicators of student progress that should be formally tested in Phase II. Of the teachers who saw measurable growth, these indicators included:

1) One teacher said student writing is the best they have seen in their three years teaching this subject.

2) Another teacher observed students asking more deliberate, specific student questions and/or fewer questions altogether.

3) A third teacher noticed that students improved structural organization on handwritten assignments, where they did not have TrueMark's assistance.

4) Another teacher found that students have stopped asking the AI Copilot questions they previously would have asked.

> *"They don't complain as much. It's just become more routine without question. It's embedded in their brain: if I'm writing an essay, I'm gonna have an intro, a minimum of two body paragraphs, and a conclusion. I stopped giving them checklists. They don't even need them."*

> *"There are very specific students that from the beginning of the year till now, they really have picked up on writing a successful paragraph or an entire essay. And I notice a change in their grades."*

> *"On the last written test I gave them, they were all very organized — more than normal. And I do wonder if that has something to do with [TrueMark]."*

Effectiveness Signal 3: Plagiarism Detection

Overall, TrueMark's positioning as a plagiarism detection tool was the platform's most important feature across all eight interviews. Many teachers referenced initial hesitancy about AI in the classroom and potential negative impacts on learning, but simultaneously recognized that the presence of AI in student life is inevitable. Reflecting on how TrueMark challenged these fears, teachers consistently framed TrueMark as such: TrueMark does not pretend AI does not exist, but ensures student work is authentically the student's. Even in cases where academic dishonesty does occur, "TrueMark catches [students] every time." Across all interviews, teachers cited this positioning as the primary reason they adopted the platform and the main priority they want to protect as TrueMark scales.

> *"I've had kids who've done three assignments, and then on the third one they use AI, and they get caught."*

> *"The work stays the student's work."*

(Unexpected) Effectiveness Signal 4: TrueMark supports teachers in identifying SEL and mental health strengths and warnings.

An unexpected, novel use-case emerged, in that 3 out of 8 teachers used TrueMark's AI Copilot engagement, rubric-checking, and edit count and timing trackers as SEL and mental health early warning signals.

Teachers reported using AI Copilot engagement patterns as a differentiation signal between motivated and disengaged students, highlighting which students asked thoughtful, content-related

questions and which ignored the AI Copilot's presence entirely. These distinctions in active learning and participation were more difficult to identify in assignments pre-TrueMark, and the platform now enables teachers to more effectively target support to students with lower levels of engagement.

As increasing numbers of students combat anxiety and mental health conditions both inside and outside of the classroom, TrueMark also appears to serve as a critical indicator of where these challenges emerge and which students may benefit from more support (CDCP, 2024). By alerting teachers to the number of times that students check their work against the rubric, TrueMark helps teachers identify specific students who might be demonstrating compulsive habits. Similarly, the edit count and and timing tracking features flag students who are working on assignments in the middle of the night or making excessive revisions to their work, signaling to teachers which students might be overly stressed or obsessive about assignments. With these data points, teachers can make informed decisions about which students to recommend to counselors.

This finding warrants attention in Phase II. If TrueMark engagement patterns can surface behavioral signals relevant to student wellbeing, there are ethical implications to explore, such as: What data should be surfaced to teachers, to counselors, and to students themselves? How should it be framed? What is TrueMark's responsibility?

Teacher Perceptions of Phase II Dashboards

In anticipation of the Phase II proposal, the research team embedded an interactive mockup of the TrueMark skills dashboard, which was populated with real student data from the pilot, into the teacher interview protocol. Researchers guided teachers directly to the dashboard during their interview session to gauge its strengths and weaknesses and to gather formative input for the Phase II build-out.

Overall, the dashboard's value proposition resonates with teachers. Two prominent themes emerged from the interviews: first, teachers expressed a desire to "drill down" into individual skills to see the specific learning objectives mapped to each and the progress on those learning objectives over time; and second, teachers wanted transparency into how the platform calculates skills levels. These findings suggest that teachers need interpretive transparency to support instructional decision-making. However, conceptually the dashboard is perceived as highly valuable, and in some cases, more valuable than nationally normed benchmark data because it provides real-time, continuous data.

> *"I think if I did this again next year, I would try to use it probably earlier in the year — almost like a benchmark. All right, they start here. Where are they going to end?"*

Teachers also requested more transparency into how the dashboard is estimating proficiency level. Additionally, the extent to which the dashboard depends on the accuracy of auto-grading with or without teacher overrides was a concern.

> *"I noticed there was some disconnect between the dashboard and the rubric checker and the chatbot… the chatbot and the rubric had told them they hadn't provided enough evidence and had given them 0 out of 3 or 0 out of 2. And then the dashboard marked them proficient in it."*

This finding has direct implications for Phase II product development. Teacher trust in the dashboard's output depends on data integrity. Resolving the data pipeline between auto-grading accuracy and phantom errors, teacher-corrected rubric grades, and skills-level calculations is a prerequisite for the dashboard to influence instructional decision-making at scale.

Key Insights for Phase II

Taken together, the pilot findings provide a promising signal of TrueMark's usability, feasibility, and early effectiveness, while identifying priority areas for refinement and development in Phase II.

Areas to refine include:

1) *AI Copilot grade level scaffolding and helpfulness.* The AI Copilot's Socratic design is well-suited for developing writers but requires grade-level differentiation; teachers reported that the AI Copilot provides more support than advanced students should receive, reducing the productive struggle that characterizes strong writing instruction at those levels.

2) *Auto-grading accuracy.* Rubric auto-grading accuracy was a concern across all eight teacher interviews, with a consistent pattern: the rubric graded too generously at upper grade levels, often awarding credit for keyword presence rather than substantive elaboration, and too strictly at lower grade levels, flagging grammar conventions not yet formally taught. Because skill estimates are derived from rubric scores, improving auto-grading accuracy is foundational to the reliability of the skills dashboard and a prerequisite for teacher trust in the platform's assessment data.

Areas for further development include:

1) *Dashboard discoverability and interpretive transparency.* The skills dashboard requires greater visibility within the platform workflow and in-platform explanations of how skill levels are calculated, what each level means, and what a student needs to do to advance. Teachers suggest in-app feature prompts and onboarding videos as concrete solutions.

2) *Dashboard skills estimate feature.* Moving into Phase II as the scoring feature is refined, TrueMark should monitor teacher overrides to ensure the platform's skill assessments accurately reflect what students know and can do, and to help the platform learn from teacher expertise over time.

Two additional findings warrant further exploration in Phase II: the potential of TrueMark engagement data as a student wellbeing signal, and the need for accessibility features (speech-to-text, dark mode, page break controls) to support a more diverse student population.

References

Brooke, J. (2013). *SUS: A retrospective*. Journal of Usability Studies, 8(2), 29–40. https://uxpajournal.org/wp-content/uploads/sites/7/pdf/JUS_Brooke_February_2013.pdf

Bruning, R., Dempsey, M., Kauffman, D. F., McKim, C., & Zumbrunn, S. (2013). Examining dimensions of self-efficacy for writing. *Journal of Educational Psychology, 105*(1), 25–38. https://doi.org/10.1037/a0029692

Centers for Disease Control and Prevention. (2024). *Youth Risk Behavior Survey Data Summary & Trends Report: 2013–2023*. U.S. Department of Health and Human Services. https://www.cdc.gov/yrbs/dstr/pdf/YRBS-2023-Data-Summary-Trend-Report.pdf

Chiu, T.K.F., Çoban, M., Sanusi, I.T., & Ayanwale, M.A. (2025). Validating student AI competency self-efficacy (SAICS) scale and its framework. *Education Tech Research Dev 73*, 2785–2807. https://doi.org/10.1007/s11423-025-10512-y

Delić, H. & Bećirović, S. (2016). Socratic Method as an Approach to Teaching. *European Researcher, 111*(10), 511-517. https://erjournal.ru/journals_n/1478004367.pdf

Ericsson, K. Anders, Ralf T. Krampe, and Clemens Tesch-Römer. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review, 100*(3), 363-406. https://doi.org/10.1037/0033-295X.100.3.363

Fabio, R. A., Plebe, A., & Suriano, R. (2024). AI-based chatbot interactions and critical thinking skills: An exploratory study. *Current Psychology, 44*(9), 8082–8095. https://doi.org/10.1007/s12144-024-06795-8

Guest, G., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis*. SAGE Publications. https://doi.org/10.4135/9781483384436

Ho, YR, Chen, BY, & Li, CM. (2023). Thinking more wisely: using the Socratic method to develop critical thinking skills amongst healthcare students. *BMC Medical Education, 23*(173), 1-16. https://doi.org/10.1186/s12909-023-04134-2

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence, 2*, 1-12. https://doi.org/10.1016/j.caeai.2021.100017

Kehrer, P., Kelly, K., & Heffernan, N. (2013). *Does immediate feedback while doing homework improve learning?* Paper presented at the 26th International Florida Artificial Intelligence Research Society Conference. https://eric.ed.gov/?id=ED615325