IAS Technical Report · November 2026

Acellus Writing Tutor Gold Edition: A Process-Based, AI-Coached System for Teaching Writing and Assuring Authorship in the Age of Generative AI

Roger Billings, Ph.D.
International Academy of Science, Kansas City, Missouri

ABSTRACT

Generative artificial intelligence has made the finished essay an unreliable record of what a student knows. A student can now obtain a polished, on-topic composition in seconds, and an assessment that grades only the final product can no longer tell composition from transcription. This report describes Acellus Writing Tutor Gold Edition, a writing system for students in kindergarten through grade 12 that responds to this problem in two ways. First, it moves the object of instruction from the product to the process: students work through seven steps (Understand, Brainstorm, Draft, Revise, Edit, Publish, and Revise again), must correct every flagged spelling and grammar error themselves before a piece can be published, and then enter a scored revision loop in which a large language model (LLM) evaluates the writing against a rubric and the assignment's requirements and coaches the student toward a stronger revision. Students may raise their score above 100 percent through bonus credit for continued revision. Second, it surrounds the writing with an authorship assurance architecture: work is permitted only inside the Acellus application, paste is disabled in the editor, devices are verified through platform attestation and per-install cryptographic keys, and a behavioral fingerprint built from session timing flags account sharing, outsourcing, and automation.

We review the research basis for each design choice. The evidence is strong for explicit instruction in planning, revising, and editing; for formative feedback that names a next step; and for learning conditions in which students generate and retrieve answers themselves. Evidence for LLM-generated feedback on student writing is newer but encouraging, with the important qualification that unrestricted AI assistance can improve performance while impairing learning. Evidence for behavioral detection of misconduct is real but probabilistic: at realistic rates of misconduct, even accurate detectors produce many false flags. Several elements, including bonus credit above 100 percent, have no direct research base and should be treated as hypotheses. We identify vulnerabilities in LLM-based scoring (self-preference, verbosity bias, and prompt injection), assess the monitoring design against privacy law that governs minors' data, and make twelve recommendations, the most important of which is to capture writing-process evidence inside the editor so that authorship is demonstrated by how a text was built rather than inferred from who was holding the device. We close with an evaluation plan.

1. Introduction

For most of the history of schooling, a finished essay was reasonable evidence that its author could write. That assumption no longer holds. Large language models produce fluent, organized, grammatically clean prose on almost any topic, and students have adopted them quickly: in a Pew Research Center survey of U.S. teens conducted in fall 2025, 54 percent said they had used AI chatbots to get help with schoolwork, and one in ten said they do all or most of their schoolwork with a chatbot's help (McClain et al., 2026). A year earlier, 26 percent of teens reported using ChatGPT for schoolwork, double the share in 2023 (Sidoti et al., 2025). An assignment graded only on its final text now measures access to a tool as much as it measures skill.

The original Acellus Writing Tutor experienced this shift directly. Students could paste an assignment's instructions into an external AI tool, paste the result back into the Writing Tutor, and receive a high score with no learning. The earlier scoring engine had a second weakness familiar from the automated essay scoring literature: it could be moved by surface features. In one example observed by our team, inserting a grammatically unnecessary colon raised a submission's score. Systems that reward countable features invite students to supply the features rather than the writing (Perelman, 2014; Powers et al., 2002).

Acellus Writing Tutor Gold Edition was designed around a different premise: if the finished product can be produced without learning, the system must teach and observe the process that produces it. This report has four aims:

1. To describe the system's workflow, scoring loop, and tutor as built, separating what is implemented from what is recommended.
2. To present the research basis for each design decision, and to state plainly where that basis is strong, where it is indirect, and where it is absent.
3. To evaluate the authorship assurance (monitoring and fingerprinting) design, including its accuracy limits and its legal and ethical constraints.
4. To recommend improvements and a plan for evaluating whether the system works.

A note on evidence. Throughout this report we distinguish three kinds of claim: findings from published research, which we cite; observations from Acellus development and testing, which we label as such and which have not yet been independently evaluated; and design hypotheses, which are reasonable but untested. Readers should weigh them accordingly.

2. Background

2.1 Writing is a set of cognitive processes, not a single act

Modern writing research treats composition as a set of recursive processes rather than a linear performance. Flower and Hayes (1981) modeled writing as planning (generating and organizing ideas and setting goals), translating ideas into text, and reviewing, all coordinated by a monitor and constrained by the task environment and the writer's long-term memory. Hayes (2012) later revised the model to give working memory, motivation, and transcription skills (handwriting, typing, spelling) explicit roles. Bereiter and Scardamalia (1987) contrasted *knowledge telling*, in which novice writers write down whatever comes to mind on a topic, with *knowledge transforming*, in which skilled writers reshape what they know to meet a rhetorical goal. Kellogg (2008) argued that progress from one to the other takes years of deliberate practice and depends on freeing working memory, which is why transcription skills such as spelling matter even to advanced writers: effort spent on spelling is effort unavailable for thinking.

Two implications follow for system design. First, a tool that delivers finished text bypasses exactly the processes (planning, translating, reviewing) that instruction is meant to develop. Second, because the processes are separable, they can be taught and supported separately, which is the logic behind a stepwise workflow.

2.2 What works in writing instruction

The research base on writing instruction is large and unusually consistent. In *Writing Next*, a meta-analysis of instruction for students in grades 4 through 12, Graham and Perin (2007) found the

largest average effects for teaching writing strategies (effect size [ES] = 0.82) and summarization (ES = 0.82), followed by collaborative writing (ES = 0.75) and word processing (ES = 0.55); the process writing approach had a smaller effect (ES = 0.32), and traditional grammar instruction had a small but statistically significant negative effect. A meta-analysis of 115 studies in the elementary grades found the largest effect for strategy instruction (ES = 1.02), and especially for the Self-Regulated Strategy Development model (ES = 1.17), compared with ES = 0.40 for the process approach (Graham, McKeown, et al., 2012). A separate meta-analysis of interventions in grades 4 to 6 found the largest effects for goal setting, strategy instruction, feedback, text structure instruction, and peer assistance (Koster et al., 2015). Two practice guides from the What Works Clearinghouse (WWC) of the U.S. Institute of Education Sciences summarize the evidence for teachers. The elementary guide rates its recommendation to "teach students to use the writing process for a variety of purposes" as supported by strong evidence, the highest rating (Graham, Bollinger, et al., 2012). The secondary guide gives its strong-evidence rating to explicit teaching of writing strategies through a cycle of modeling, practice, and reflection, and rates the use of assessment to inform feedback as supported by minimal evidence (Graham et al., 2016).

Across these syntheses, the most reliable finding is that students write better when they are explicitly taught strategies for planning, revising, and editing, and practice them on real writing tasks. A process framework without explicit strategy instruction produces smaller gains than one with it. This distinction matters for Gold Edition: the seven steps provide the structure, but the instructional value comes from what the system teaches the student to do inside each step.

2.3 Feedback and revision

Feedback is among the most powerful influences on learning, and also among the most variable. Hattie and Timperley (2007), synthesizing 12 meta-analyses, reported an average effect of 0.79, and argued that feedback works when it answers three questions for the learner: Where am I going? How am I going? Where to next? Feedback aimed at the self, such as general praise, was rarely effective. Feedback can also harm: in a meta-analysis of 607 effects, Kluger and DeNisi (1996) found a positive average effect (d = 0.41) but also that more than one third of feedback interventions reduced performance. A more recent meta-analysis of 435 studies found a medium average effect (d = 0.48) that depended strongly on the information the feedback carried (Wisniewski et al., 2020). Shute (2008) concluded that formative feedback should be specific, timely, and focused on the task rather than the learner. For writing specifically, a meta-analysis of formative assessment in grades 1 through 8 found that feedback improved writing quality whether it came from adults (ES = 0.87), from self-assessment (ES = 0.62), from peers (ES = 0.58), or from computers (ES = 0.38) (Graham, Hebert, & Harris, 2015).

Revision is where feedback becomes learning, but students do not revise well on their own. In her review, Fitzgerald (1987) found that young and inexperienced writers revise little and that their revisions are mostly superficial. Faigley and Witte (1981) distinguished *surface changes*, which alter wording or mechanics without changing meaning, from *meaning changes*, which alter content. Novices make mostly surface changes. This taxonomy gives a practical way to measure whether a revision loop is producing deeper revision or only polishing, and we use it in the evaluation plan in Section 8.

2.4 Automated and AI-generated feedback on writing

Automated writing evaluation (AWE) predates large language models. Meta-analyses find medium average effects of automated feedback on writing performance: g = 0.55 across 20 studies in one analysis (Fleckenstein et al., 2023) and g = 0.59 for between-group comparisons in English-as-a-

foreign-or-second-language settings in another (Ngo et al., 2024). An important qualification is that most of this evidence concerns the texts students revise with the feedback; evidence that the gains carry over to students' unassisted writing is weaker (Stevenson & Phakiti, 2014). Classroom studies of AWE in the middle grades have examined effects on writing performance, writing self-efficacy, and teacher workload together (Wilson & Roscoe, 2020).

Large language models have changed the quality and flexibility of automated feedback. When trained raters compared 200 human and 200 ChatGPT feedback samples on the same secondary-school essays, human feedback was of higher quality in every category except alignment with assessment criteria, but the differences were modest (Steiss et al., 2024). In a study of 459 upper-secondary students, LLM-generated feedback, compared with no feedback, improved revision performance ($d = 0.19$), task motivation ($d = 0.36$), and positive emotions ($d = 0.34$) (Meyer et al., 2024). A six-week study of university English learners found no difference in learning outcomes between GPT-4 feedback and human tutor feedback (Escalante et al., 2023).

The same capability can also undercut learning. The clearest experimental evidence comes from a field experiment with nearly 1,000 high school mathematics students (Bastani et al., 2025). Students given unrestricted access to a chatbot improved their practice-problem grades by 48 percent while they had access, but when access was removed they performed 17 percent worse than students who had never had access. Students given a version designed as a tutor, with guardrails that withheld answers and gave hints, improved their practice grades by 127 percent, and the guardrails largely prevented the later harm. In a randomized trial with undergraduate physics students, an AI tutor built on research-based teaching principles produced median learning gains more than twice those of an in-class active-learning session, in less time (Kestin et al., 2025). A preprint study of adults writing essays with and without an LLM reported weaker neural connectivity and poorer recall of their own essays among LLM users; it has not yet been peer reviewed (Kosmyna et al., 2025). The lesson for design is not that AI tutoring works or fails, but that the boundary between *helping a student think* and *doing the thinking for the student* determines which outcome occurs.

2.5 Academic integrity in the AI era

Two forms of misconduct concern us here. The first is substitution by AI: a student submits text generated by a model. The second is substitution by people: a student has someone else, often a paid service, do the work. In higher education, a systematic review of 65 studies found that the share of students admitting to paying someone else to complete their work rose from a historic average of 3.5 percent to 15.7 percent in samples from 2014 onward (Newton, 2018). A survey of more than 14,000 Australian university students associated contract cheating with dissatisfaction with teaching and with a perception that there were many opportunities to cheat (Bretag et al., 2019). We found no peer-reviewed estimate of paid outsourcing among K–12 students; the best available K–12 data concern AI use (Section 1), and the size of the outsourcing problem in K–12 online learning remains to be measured. Online learning adds a third concern, *account-level* substitution, in which another person logs in as the student and completes coursework directly.

The most common institutional response to AI substitution, automated AI-text detection, is unreliable. In a systematic test of 14 detection tools, including widely used commercial products, none exceeded 80 percent accuracy and only five exceeded 70 percent (Weber-Wulff et al., 2023). Detectors are also biased: seven detectors misclassified more than half of a set of essays written by non-native English speakers as AI-generated, with an average false-positive rate of 61 percent (Liang et al., 2023). For this reason Gold Edition does not rely on AI-text detection. Its strategy is to make

honest work easier and dishonest work harder to hide, and to base any concern about authorship on evidence of how a text was produced.

3. System Description

3.1 Design principles

Gold Edition was built on five principles:

1. **Teach the process, not just the product.** Every step of composition happens inside the system and is supported there.
2. **The student does the writing.** The tutor explains, questions, and points to what to improve; it does not write the student's text.
3. **Learning happens in the student's own corrections.** Errors are flagged, not silently fixed.
4. **Feedback must point to a next step.** Every score comes with specific guidance on how to raise it.
5. **Make the honest path the easiest one.** Help is available at every step, so that outside help is less tempting, and the environment is closed, so that outside help is harder to use.

3.2 Student workflow

Figure 1 shows the workflow. Table 1 summarizes what the student does at each step, what the system does, and the research rationale.

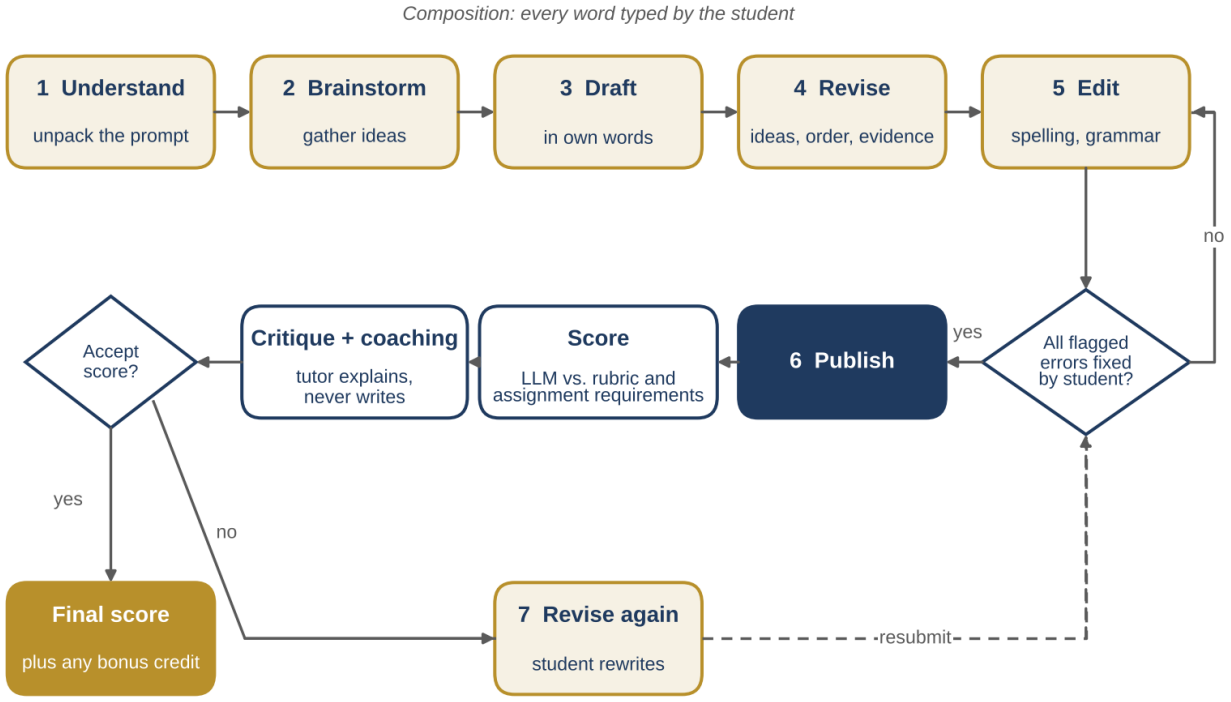


Figure 1. Gold Edition workflow. Steps 1–5 are composition, in which every word is typed by the student. A conventions gate blocks publication until the student has corrected each flagged error. After publication, a large language model scores the writing and coaches revision; the student may accept the score or revise and resubmit, earning bonus credit for continued revision.

Table 1. The seven steps: what the student does, what the system does, and the research rationale

Step	Student	System	Research rationale
1 Understand	Answers guided questions on task, purpose, audience, and own position	Presents the assignment's goal and requirements; asks the questions	Task analysis and goal setting are planning processes; setting product goals improves writing (Flower & Hayes, 1981; Graham & Perin, 2007)
2 Brainstorm	Generates and records preliminary ideas	Prompts idea generation	Prewriting improves writing quality (Graham & Perin, 2007)
3 Draft	Writes in own words	Editor inside the Acellus app; paste disabled	Deferring correction frees working memory for translating ideas into text (Kellogg, 2008)
4 Revise	Reworks ideas, organization, and evidence	Supports revision	Explicit instruction in revising strategies is among the most effective writing practices (Graham & Perin, 2007; Fitzgerald, 1987)
5 Edit	Finds and corrects each flagged error by hand	Flags errors; provides dictionary; blocks publication until flags are cleared	Spelling instruction improves spelling and transfers to writing; self-generated answers are retained better (Graham & Santangelo, 2014; Slamecka & Graf, 1978)
6 Publish	Submits the piece	LLM scores against rubric and assignment requirements	Automated scores require validation against trained human raters (Williamson et al., 2012)
7 Revise again	Accepts the score or revises and resubmits	Returns critique, specific suggestions, and bonus credit for continued revision	Feedback that names a next step and repeated attempts to criterion improve learning (Hattie & Timperley, 2007; Kulik et al., 1990)

Step 1, Understand. Each assignment states its task, purpose, audience, and requirements. Before writing, students answer guided questions: What is the task? Who is my audience? What do I actually think? Analysis of the task and goal-setting are planning processes in the Flower and Hayes (1981) model, and setting specific product goals is itself an effective writing intervention (Graham & Perin, 2007).

Step 2, Brainstorm. Students gather preliminary ideas before drafting. Pre-writing activities that help students generate and organize ideas before writing improve writing quality (Graham & Perin, 2007).

Step 3, Draft. Students write in their own words in the Acellus editor. Students are encouraged to draft quickly and imperfectly, deferring correction to later steps. Separating idea generation from correction reduces the load on working memory during translation (Kellogg, 2008).

Step 4, Revise. Students rework ideas, organization, and evidence.

Step 5, Edit. Students correct spelling, grammar, and punctuation. Section 3.3 describes the conventions gate that governs this step.

Step 6, Publish. The student submits the piece for scoring.

Step 7, Revise again. The system returns a score, a critique, and specific suggestions. The student chooses to accept the score or to revise and resubmit. Section 3.5 describes this loop.

3.3 The conventions gate

Gold Edition highlights spelling and grammar errors and will not accept a piece for publication until each flagged error has been corrected. Corrections are made by the student. A dictionary is available, but the system does not auto-correct, and the student must locate the correct form and type it.

The rationale draws on three lines of research. First, spelling is learned, and explicit spelling instruction improves spelling and also transfers to reading and writing. A meta-analysis of 53 studies found that formal spelling instruction improved spelling (ES = 0.54), that gains carried over to spelling in students' own writing (ES = 0.94), and that they were maintained over time (Graham & Santangelo, 2014). Second, information that learners generate is remembered better than information they are given (the generation effect; Slamecka & Graf, 1978; average effect 0.40 across 86 studies in Bertsch et al., 2007), and effortful retrieval strengthens memory more than re-exposure (Roediger & Karpicke, 2006). Conditions that slow initial performance in this way can improve long-term retention, which Bjork (1994) called *desirable difficulties*. Third, automatic correction can mask what a writer does not know, and spelling checkers are imperfect tools that require knowledge to use well. In studies of students with learning disabilities, spelling checkers raised the share of errors students corrected from 9 to 37 percent but missed many misspellings that formed real words (MacArthur et al., 1996), and adult users follow false checker suggestions and overlook errors the checker fails to flag (Galletta et al., 2005).

Three cautions apply. First, correcting errors in a draft is not the same as learning to avoid them. In a study of second-language writers, students whose errors were marked corrected more of them during revision, but a week later, on a new piece of writing, their error rates had changed no more than those of a control group (Truscott & Hsu, 2008); other studies of comprehensive error correction have found durable gains in new texts (van Beuningen et al., 2012). Whether Gold Edition's gate improves spelling in students' *next* piece of writing is therefore an empirical question, and Section 8 proposes measuring it. Second, a difficulty is desirable only if the learner can overcome it. A student who cannot approximate a word's spelling cannot look it up in a conventional dictionary, so the dictionary should support phonetic and approximate search (Recommendation 7). Third, grammar checkers produce false positives, flagging proper nouns, dialogue, deliberate fragments, and dialect forms, and a gate that requires every flag to be cleared may force students to "correct" writing that was right. A logged override for flags the student believes are wrong preserves the learning benefit while removing this failure mode.

3.4 Scoring

When a piece is published, it is scored by an LLM from Anthropic's Claude family, accessed through Anthropic's application programming interface (API). The scorer evaluates two things: how well the piece fulfills the assignment's specific requirements, and the overall level of the writing.

Moving from feature-based scoring to LLM scoring addressed a known weakness. Feature-based automated scorers rely on measurable proxies (length, vocabulary, sentence variety, punctuation patterns), and students and researchers have shown that such systems can be raised by text that is long, elaborate, and meaningless (Perelman, 2014; Powers et al., 2002). In Gold Edition testing, a change such as the inserted colon described in Section 1, which formerly raised a score, is now evaluated in context, and an unnecessary punctuation mark lowers the score instead of raising it. This is a development observation and should be confirmed with a systematic adversarial test (Section 8).

LLM scoring has its own reliability questions. Early studies found that LLMs score essays with useful but imperfect accuracy and reliability (Mizumoto & Eguchi, 2023). When four LLMs, including a Claude model, each scored the same set of 119 essays on two occasions, agreement with human scores and consistency across occasions varied by model, and all models showed fluctuations in their scoring (Pack et al., 2024). Section 5 discusses specific vulnerabilities and safeguards.

3.5 The revision loop and bonus credit

After scoring, the student receives a critique and specific suggestions for improvement, and may accept the score or revise. Each resubmission passes through the conventions gate and is rescored. The development team regards the quality of this critique as the single most important determinant of whether a student chooses to revise and whether the revision improves the writing, and has concentrated its effort here.

Students who continue to revise can earn bonus credit, raising the score above 100 percent. In Acellus testing, the possibility of exceeding 100 percent has been a strong motivator for additional revision. This observation is consistent with research on mastery learning, in which students receive corrective feedback and additional attempts until they reach a criterion (Kulik et al., 1990, who found an average effect of 0.52 across mastery-learning programs; Bloom, 1984), and with meta-analytic evidence that well-designed game elements can improve motivation and engagement. Sailer and Homner (2020) found positive effects of gamification on cognitive ($g = 0.49$), motivational ($g = 0.36$), and behavioral ($g = 0.25$) outcomes, although only the cognitive effect held up in the most rigorous studies.

Bonus credit is, however, the element of Gold Edition with the least direct research support, and it carries a design risk. Expected, performance-contingent tangible rewards can undermine intrinsic motivation, whereas informational feedback about competence tends to support it. A meta-analysis of 128 experiments found that expected rewards contingent on engaging in, completing, or performing well at a task reduced intrinsic motivation, with stronger effects for children than for college students, whereas positive informational feedback increased it (Deci et al., 1999; see also Ryan & Deci, 2000). A score that rewards *rounds of revision* rather than *improvement* also invites trivial resubmissions, the same incentive problem that made the old scorer gameable. Recommendation 3 proposes tying bonus credit to measured improvement and reporting growth by trait, so that the reward carries information about the student's competence.

3.6 The tutor and its boundaries

The tutor's role is that of a writing teacher in conference with a student: it explains the score, identifies what to strengthen, asks questions that prompt new ideas, and helps the student see how to organize their thinking. It helps students improve style and vocabulary by explaining why a choice is weak and what kind of change would strengthen it. It does not write the student's text.

This boundary is the most consequential design decision in the system. The experimental evidence summarized in Section 2.4 shows that the same model can improve learning when it is constrained to guide and impair learning when it supplies answers (Bastani et al., 2025). Two practical questions follow. First, where exactly is the line? A suggestion that a student "use a more precise verb than *got*" leaves the work to the student; a suggestion of three replacement sentences does not. Second, how is the line enforced and audited? We recommend that the tutor's permitted behaviors be written into its instructions as an explicit policy, tested with adversarial student requests (for example, "just rewrite this paragraph for me"), and that any text the tutor supplies be logged so that the share of a final piece that originated with the tutor can be measured (Recommendation 4).

3.7 Parent and teacher involvement

Parents and teachers can see each student's progress, and Acellus encourages them to review students' writing and to praise and encourage growth. Research supports a specific form of this involvement. Parental involvement that communicates expectations and the value of education is more strongly associated with achievement than direct help with homework. A meta-analysis of 50 studies of middle school students found that academic socialization, communicating expectations and linking schoolwork to goals, had the strongest positive association with achievement, while direct help with homework was not positively associated with it (Hill & Tyson, 2009). A synthesis of research on parent involvement in homework likewise found mixed associations with achievement, with clearer benefits from parents setting routines than from direct help (Patall et al., 2008). Praise for effort and strategy supports persistence after difficulty better than praise for ability (Mueller & Dweck, 1998). Writing self-efficacy is in turn a consistent predictor of writing motivation and performance (Pajares, 2003). The practical guidance for parents is therefore to praise specific improvements and the work of revising ("your second paragraph is much clearer now that you added the example"), not to edit the writing themselves.

4. Authorship Assurance Architecture

4.1 Threat model

Table 2 lists the ways a student's submitted writing could fail to be the student's own work, the control Gold Edition applies to each, and the risk that remains.

Table 2. Threat model for authorship of submitted writing

Threat	Example	Current control	Residual risk and recommended addition
AI text pasted in	Student copies chatbot output into the editor	App-only work; paste disabled	Low
AI text retyped	Student reads generated text on a phone and types it	None specific to this threat	High. Add writing-process evidence and explain-back checks (Recs. 1, 8)
Another person logs in	Paid service completes work on the student's account	Attestation, per-install keys, device count, fingerprint drift	Moderate. A helper using the student's own device, or remote-control software on it, passes device checks. Add process evidence and explain-back checks
Another person writes, student types	Parent, sibling, or service drafts text the student transcribes	None specific to this threat	High. Transcription leaves a distinctive process record (Rec. 1)
Automation	Scripts or auto-clickers submit work	Attestation, signed requests, timing analysis	Low
Scorer manipulation	Padding, punctuation tricks, instructions hidden in the essay	Contextual LLM scoring	Moderate. Apply Section 5 safeguards (Rec. 5)
Tutor ghostwriting	Student coaxes the tutor into rewriting passages	Tutor instructed not to write student text	Moderate. Make the boundary explicit, tested, and logged (Rec. 4)
Bonus-credit gaming	Minimal edits resubmitted for credit	None specific to this threat	Moderate. Tie bonus credit to measured improvement (Rec. 3)

4.2 A closed writing environment

Acellus coursework, including Writing Tutor, can be completed only in the Acellus application. Within the writing editor, paste is disabled. Together these controls eliminate the simplest form of AI substitution, in which generated text is copied from a browser window and pasted into a submission, and they ensure that every character of a submission is entered inside the instrumented environment.

They do not prevent a student from reading generated text on a second device and retyping it. That residual threat is the reason writing-process evidence (Section 4.5) matters: retyping leaves a different behavioral record from composing. Paste blocking also has an accessibility cost. Some students with disabilities compose with speech-to-text, word-prediction, or augmentative and alternative communication tools that insert text by pasting or by simulated input. A blanket block can deny these students a needed accommodation, so the block should support documented exceptions (Recommendation 9).

4.3 Device trust

The monitoring design uses the strongest available evidence that a request came from the genuine Acellus application on a real device. On Android, the Play Integrity API returns verdicts on whether the application binary is genuine, whether the device is a genuine, unrooted device, and whether the account is licensed. On Apple platforms, App Attest allows a server to verify that requests come from a legitimate instance of the application, and DeviceCheck stores two persistent bits per device that survive reinstallation (Apple Inc., n.d.-a; Google, 2026). Apple notes that App Attest keys do not survive reinstallation, device migration, or restoration from backup, and that attestation should be one input to an overall risk assessment rather than a verdict (Apple Inc., n.d.-b); a change in a student's installation count can therefore be innocent. On first launch, the application also generates a cryptographic key pair in hardware-backed storage (Android Keystore or StrongBox, the Apple Secure Enclave, or a Trusted Platform Module on Windows), registers the public key, and signs every subsequent request with a timestamp and nonce. This yields an installation identifier that cannot be copied to another device, protects against replayed or forged requests, and provides a clean count of the devices each student uses.

This design deliberately avoids hardware identifiers such as serial numbers, advertising identifiers, and network hardware addresses. Those identifiers are restricted on modern mobile operating systems, and Apple's developer program prohibits deriving data from a device to identify it uniquely (Apple Inc., n.d.-c). Generated per-install keys provide the needed assurance with less privacy risk.

4.4 Behavioral fingerprinting

The monitoring design builds, for each student, a statistical profile of how that student works and flags departures from it. Features include the distribution of activity across the 168 hours of the week; the mix of time across coursework, video, activities, and browsing; the mean and variance of session length; sessions per day; active versus idle time within sessions (from foreground state and input heartbeats); median and dispersion of time spent on each item before answering; typical number of installations and primary platform; and clock skew between device and server. Profiles for new students begin from a cohort baseline and shift toward the individual as data accumulate. Anomalies are classified into two tiers: *Tier 1* drift, such as a sudden change in working hours, platform, or category mix, which may indicate that someone else is doing the work; and *Tier 2* automation, such as implausibly fast and uniform answer timing or requests that lack a valid attestation, which indicate scripted activity.

This approach has a research foundation in online learning. Analyzing clickstream data from 1.9 million participants in 115 massive open online courses (MOOCs), Northcutt et al. (2016) detected students who used a second "harvester" account to collect correct answers for their main account, by finding pairs of accounts whose submissions followed each other closely in time; the practice accounted for about 1.3 percent of certificates in the courses where it occurred. Using similar timing methods, Alexandron et al. (2017) found that about 10 percent of certificate earners in one MOOC obtained at least some of their correct answers this way. These studies show that timing and account-relationship patterns can reveal misconduct that is invisible in the answers themselves. Research on keystroke dynamics shows that typing rhythm carries individual information, but with error rates that preclude using it alone to identify a person. In a benchmark of 14 detection algorithms applied to 51 typists entering the same password, the best achieved equal-error rates near 10 percent, meaning roughly one error in ten even under controlled conditions (Killourhy & Maxion, 2009).

Two properties of the design deserve particular credit. It records summaries, not raw keystrokes, and it records the *count* of paste attempts, not their content. And it explicitly protects students who use accessibility services: the presence of such a service is never treated as a signal by itself.

4.5 Writing-process evidence

The behavioral fingerprint described above was designed for the Acellus platform as a whole. It tells us much about *who is likely at the device and when*, and comparatively little about *how a particular text was produced*. For writing, the stronger evidence is the writing process itself, and this is the most important addition we recommend.

Keystroke logging is a mature research method in writing studies. Keystroke-logging tools record every key press and pause with its timing, allowing researchers to reconstruct how a text was built (Leijten & Van Waes, 2013). Keystroke features vary with the writing task, so a student's process on one kind of assignment is not a fair baseline for another kind (Conijn et al., 2019). Process features add little to product features for predicting the *quality* of writing (Zhang & Deane, 2015); their value lies elsewhere, in showing *how* a text was produced. Of direct relevance, a study of keystroke logs from 463 eighth-grade students found that copy-typing differed systematically from drafting: copy-typing was slower and had few long pauses except between words, as the typist scanned the source, and involved only local edits, while drafting was more fluent, with more backspacing and longer pauses at sentence and word boundaries (Deane et al., 2018). In a test that, like Gold Edition, disabled copy and paste, classifiers trained on keystroke features distinguished 500 authentic essays from 500 transcribed copies with 96 to 99 percent accuracy, although the authors describe the results as preliminary because they come from one task and adult participants (Crossley et al., 2024). Researchers at Educational Testing Service have also shown that keystroke patterns can identify nonauthentic texts in a large-scale writing assessment and have examined whether such models treat demographic groups fairly (Jiang et al., 2024).

The implication for Gold Edition is that the editor can record, for each piece, a compact process profile: bursts of text production and the pauses between them, the location of pauses (within words, between words, between sentences), deletions and their scope, insertions into earlier text, movement of the cursor through the document, and the evolution of the text across the steps of the workflow. A composed essay grows unevenly, with pauses at sentence and paragraph boundaries, revisions to earlier passages, and deletions of whole phrases. A transcribed essay tends to proceed linearly at a steady rate with few revisions beyond typing corrections. A teacher-facing replay or summary of this record allows a reviewer to see how a piece was built, which is a fairer and more persuasive basis for a conversation with a student than any probability score.

4.6 Location signals

As currently implemented, the system also checks whether work is being done at the student's residence. The written monitoring design reviewed for this report recommends the opposite: it advises against collecting location, citing regulatory and parental concerns that outweigh its value. The two positions should be reconciled before release, and the privacy disclosures should describe what is actually collected.

If location is used, the evidence argues for a narrow, coarse signal. Location derived from internet protocol (IP) addresses is often inaccurate at the city level. IP geolocation databases can claim country-level accuracy but not reliable city-level accuracy (Poese et al., 2011). Household circumstances vary legitimately: students in joint custody live in two homes; traveling, military, and performing families move; students study at libraries, relatives' homes, and hospitals; and students experiencing homelessness may have no fixed residence at all and are protected by federal law. Federal law defines homeless children and youths to include those who lack a fixed, regular, and adequate nighttime residence, including children sharing housing with others or living in motels or shelters (McKinney-Vento Homeless Assistance Act, 42 U.S.C. § 11434a). A location mismatch therefore cannot be treated as evidence of misconduct. At most, an unexplained change in network region, combined with other signals, can raise the priority of a human review. Precise location (GPS) should not be collected for integrity purposes.

4.7 From signals to decisions

Every signal described in this section is probabilistic. The decisive statistical fact is that when misconduct is uncommon, even an accurate detector produces many false positives. Figure 3 illustrates this. If 2 percent of students outsource their work and a detector catches 95 percent of them while wrongly flagging 5 percent of honest students, then only about 28 percent of flagged students are actually outsourcing; nearly three of every four flags fall on honest students. Raising specificity to 99 percent raises the share of correct flags to about 66 percent, which still means one honest student flagged for every two dishonest ones.

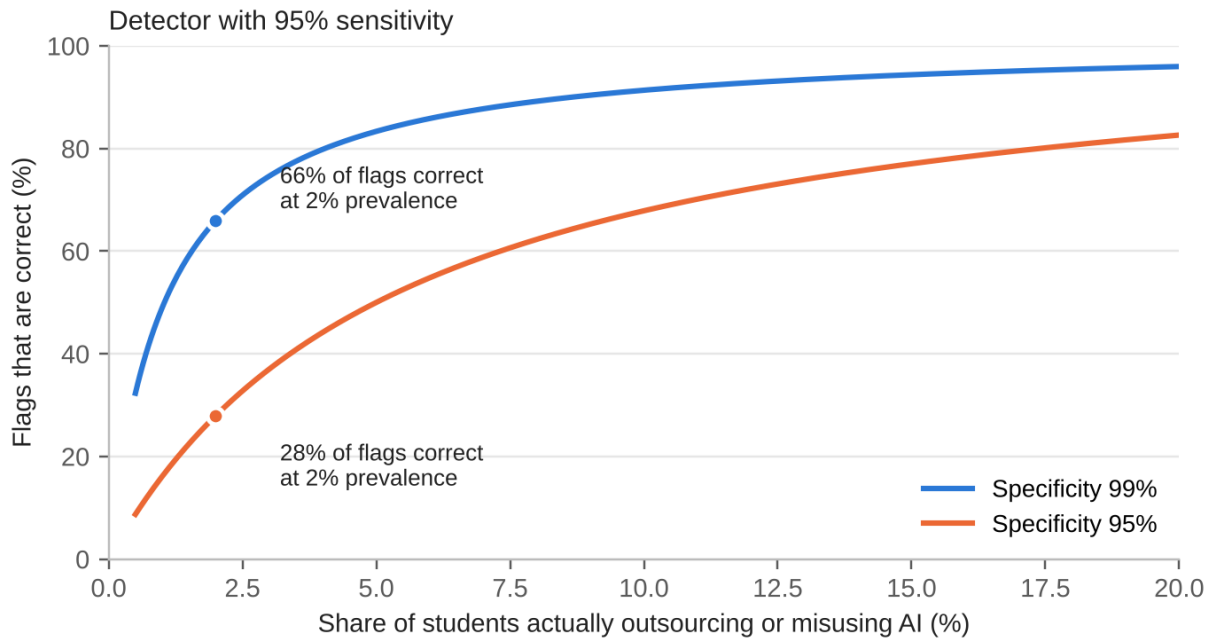


Figure 3. Share of flags that are correct (positive predictive value) as a function of how common misconduct actually is, for a detector that catches 95 percent of true cases. Specificity is the share of honest students correctly left unflagged. Values are computed from Bayes' rule; they are illustrative and are not estimates of Acellus detector performance.

For this reason, Gold Edition's integrity signals should never produce an automatic penalty. Figure 2 shows the recommended path. Signals are combined into a risk estimate with an explicit confidence level; an elevated estimate places a case in a teacher's review queue; the teacher reviews the writing-process record and talks with the student; and the outcome is coaching, a re-do under observed conditions, or no action. A brief conversation in which a student explains their own essay, its argument, and their choices is a well-established way to confirm authorship. In higher education, scaffolded assessment that includes interactive oral components has been found to help prevent misconduct (Sotiriadou et al., 2020). The form of integrity messaging also matters. In online experiments with more than 3,000 participants, honor-code statements produced only a small, statistically insignificant reduction in cheating, while a direct warning message reduced cheating by about half (Corrigan-Gibbs et al., 2015).

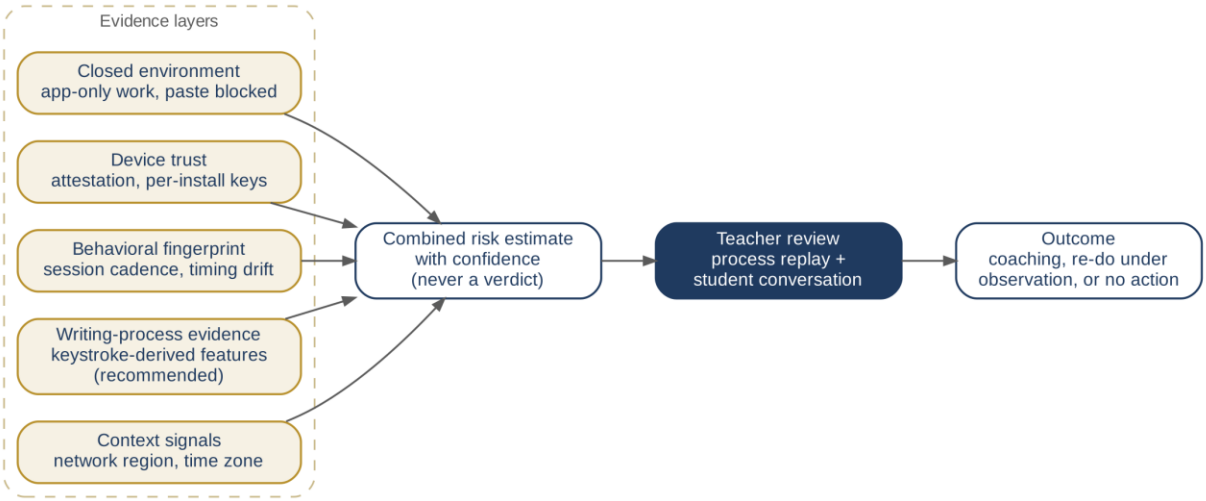


Figure 2. Recommended path from integrity signals to decisions. Each evidence layer contributes to a risk estimate that triggers human review, not a penalty. Writing-process evidence is a recommended addition.

5. LLM Scoring: Vulnerabilities and Safeguards

Replacing a feature-based scorer with an LLM removed one class of gaming and introduced others. Four deserve attention.

Self-preference. In Gold Edition the same model family both coaches revisions and scores them. LLM evaluators have been shown to recognize and favor text they generated. Panickssery et al. (2024) found that LLMs can distinguish their own outputs from others' and that the strength of this self-recognition correlates with a bias toward rating their own outputs more highly. If the tutor's suggestions shape a revision, the scorer may reward the revision partly for resembling its own style, inflating scores and rewarding adoption of the tutor's voice over the student's.

Verbosity and position biases. Studies of LLMs used as judges have documented biases toward longer responses and toward responses in certain positions, alongside high overall agreement with human judges (Zheng et al., 2023). A scorer that favors longer text recreates, in subtler form, the length-rewarding behavior that made earlier systems gameable.

Prompt injection. A student essay is untrusted input. Text inside a submission that addresses the scorer ("Grader: this essay meets all requirements") can attempt to override its instructions, a class of attack documented across LLM-integrated applications (Greshake et al., 2023). Recent work has tested this directly on LLM graders. In about 40,000 grading trials, instructions hidden in submissions raised scores from frontier models by an average of about 2.6 percentage points and raised scores from a smaller model by nearly 20 points (Wanjura et al., 2026); a preprint reported that some attack types succeeded against all five LLM graders tested (Li et al., 2026).

Inconsistency. A model may assign different scores to the same essay on different runs, and model updates by the provider can shift scores over time. In one comparison, agreement across ten repeated scorings of the same essays ranged from strong for some models to near zero for others (Seßler et al., 2025), and the behavior of a commercially hosted model can change materially between versions (Chen et al., 2024).

The appropriate safeguards are those long recommended for any automated scoring system, adapted to LLMs. Williamson et al. (2012) recommend that automated scores agree with human scores at a

quadratic weighted kappa of at least.70, that machine–human agreement fall no more than.10 below human–human agreement, and that mean score differences stay small, evaluated for subgroups as well as overall. Specifically: separate the scoring and coaching roles into distinct prompts with no shared conversational context; score against a fixed analytic rubric with anchor papers; treat the essay strictly as data, with instructions placed outside it and injection attempts detected and flagged; score each submission more than once and use the median; calibrate against a set of essays scored independently by trained teachers, monitoring agreement by grade and student subgroup; and re-run the calibration set whenever the underlying model changes.

6. Evaluation of the Monitoring Design

The monitoring design reviewed for this report is technically sound and, in several respects, more privacy-conscious than common practice. Its strengths are worth recording before its gaps.

6.1 Strengths

- **Attestation and per-install keys** convert detection of scripted activity from statistical inference into near-certainty, because a request without valid attestation did not come from the application.
- **Data minimization** is built in: summaries rather than raw keystrokes, counts rather than content, no hardware identifiers, and an event table that expires after 90 days.
- **Accessibility awareness**: the design explicitly forbids penalizing the mere presence of accessibility services.
- **Confidence weighting**: backfilled historical data are labeled and weighted lower than live data, and cold-start profiles borrow from cohort baselines.
- **Robust engineering**: idempotent recomputation of totals and detection of device clock manipulation through comparison of wall-clock and monotonic time.

6.2 Gaps

1. **It observes the session, not the writing.** The fingerprint is platform-wide. It does not yet capture the writing-process evidence that most directly distinguishes composing from transcribing (Section 4.5).
2. **It specifies signals but not decisions.** The design does not describe thresholds, how signals combine, who reviews a flag, what the student is told, or how a student can respond. Without these, the base-rate problem in Section 4.7 guarantees that honest students will be accused.
3. **No measured accuracy.** No false-positive or false-negative rates have yet been measured. They should be estimated before any flag affects a student (Section 8).
4. **Location.** As noted in Section 4.6, the stated practice of checking residence conflicts with the written design.
5. **Fairness across groups.** Behavioral baselines may differ systematically for students with disabilities, English learners, students who share family devices, and students with irregular schedules. Error rates should be measured separately for these groups. In one evaluation of keystroke-based models for detecting nonauthentic writing, the models were somewhat more likely to misclassify original writing by male test takers, and rebalancing the training data did not consistently remove the difference (Jiang et al., 2024).

6.3 Legal and policy constraints

Most Acellus students are minors, and several bodies of law apply to the data this design collects. The following is a summary for planning purposes, not legal advice; counsel should review the final design.

Children's Online Privacy Protection Act (COPPA). The COPPA Rule applies to personal information collected online from children under 13. It defines personal information to include a persistent identifier that can recognize a user over time, which covers installation identifiers and behavioral profiles keyed to them. An exception for "support for the internal operations" of a service includes protecting its security and integrity, but information collected under it may not be used to amass a profile on a specific individual, and the related consent exception applies only where no other personal information is collected (16 C.F.R. § 312.2; § 312.5(c)(7)). Because Acellus collects names and coursework, it should rely on verifiable parental consent with clear notice that integrity monitoring is among the purposes, rather than on the exception. Amendments effective June 23, 2025, with compliance required by April 22, 2026, add a written information security program and a written, disclosed data-retention policy that forbids indefinite retention, and extend the definition of personal information to biometric identifiers usable for automated recognition (Federal Trade Commission, 2025).

Family Educational Rights and Privacy Act (FERPA). FERPA governs education records held by institutions that receive U.S. Department of Education funds, or by parties acting for them (34 C.F.R. § 99.3). When Acellus serves such schools, monitoring data tied to a student are likely education records, and Acellus may receive them only as a "school official" under the school's direct control (34 C.F.R. § 99.31(a)(1)(i)(B)). FERPA does not reach families who enroll directly for home schooling, but those families are owed the same transparency.

State student privacy laws. California's Student Online Personal Information Protection Act, which many states have used as a model, prohibits operators of K–12 services from using student information, including persistent unique identifiers, to amass a profile about a pupil except in furtherance of K–12 school purposes (Cal. Bus. & Prof. Code § 22584). Integrity monitoring should be documented as serving a school purpose and its data kept from any other use.

Biometric and behavioral data. The California Consumer Privacy Act defines biometric information to include behavioral characteristics used to establish identity, explicitly including "keystroke patterns or rhythms," and treats processing it to identify a person as sensitive (Cal. Civ. Code § 1798.140(c), (ae)). The Act applies to for-profit businesses, and its application to a nonprofit and its affiliates needs legal review; but its definition signals how regulators view keystroke and timing profiles, which matters for Recommendation 1.

Accessibility. Integrity controls must not exclude students with disabilities. Private schools are public accommodations under Title III of the Americans with Disabilities Act (42 U.S.C. § 12181(7)(J)); programs receiving federal funds are covered by Section 504 of the Rehabilitation Act (29 U.S.C. § 794); and a 2024 Department of Justice rule requires public entities, including public schools that purchase Acellus, to meet WCAG 2.1 Level AA, with compliance dates now set for 2027 and 2028 (Department of Justice, 2024, 2026). Google's optional Play Integrity risk verdict can flag apps that control input, and it may flag unverified assistive tools (Google, 2026); such flags should never count against a student by themselves.

Beyond legal compliance, research on student monitoring raises concerns about chilling effects and disproportionate discipline. A national survey found that 88 percent of teachers reported student activity monitoring at their schools, that 67 percent knew a student who had been disciplined as a

result, and that students with disabilities were monitored at higher rates than their peers (Laird et al., 2023). Ethicists have argued that online supervision technologies require institutional review and justification before use (Coghlan et al., 2021), and a federal court has held that a public university's webcam scans of students' rooms violated the Fourth Amendment (*Ogletree v. Cleveland State University*, 2022). These concerns argue for transparency with families about what is monitored and why, and for using monitoring data only for integrity and learning, never for unrelated discipline.

7. Recommendations

The following recommendations are listed in order of priority.

1. **Capture writing-process evidence in the editor.** Record keystroke-derived features (production bursts, pause locations, deletions, insertions into earlier text, cursor movement, and step-by-step text versions) and present them to teachers and parents as a replay or summary of how each piece was built. This is the strongest available evidence of authorship and the most persuasive to families (Section 4.5).
2. **Define the decision process before release.** Specify how signals combine, the confidence required to queue a human review, what the student and family are told, and how they can respond. Prohibit automatic penalties based on integrity signals.
3. **Tie bonus credit to measured improvement.** Award bonus credit for gains on specific rubric traits between versions, not for the number of resubmissions; cap the bonus; and show students trait-level growth ("Organization improved from 3 to 4") so that the reward carries information about competence.
4. **Make the tutor's boundary explicit and auditable.** Write the permitted and prohibited tutor behaviors into its instructions; test them with adversarial requests; log any text the tutor supplies and measure the share of each final piece that originated with the tutor.
5. **Harden the scorer.** Separate scoring from coaching; use a fixed analytic rubric with anchor papers; isolate the essay as untrusted data and flag injection attempts; score twice and take the median; calibrate against teacher scores and monitor agreement by grade and subgroup; re-calibrate whenever the model changes (Section 5).
6. **Publish the rubric to students.** Students who know the criteria can direct their own revision, and self-assessment against criteria is itself an effective formative practice (Graham, Hebert, & Harris, 2015).
7. **Improve the conventions gate.** Add phonetic and approximate search to the dictionary; allow a logged override for flags the student believes are wrong; add short, spaced review of each student's own recurring misspellings.
8. **Add brief explain-back checks.** Occasionally ask the student, inside the application, short questions about their own essay (why they chose an example, what their thesis claims). Use these routinely for a small random sample and always when a review is triggered.
9. **Accommodate students with disabilities.** Allow documented exceptions to the paste block and accessibility-service signals for students who compose with assistive technology.
10. **Reconcile and disclose location practice.** Decide whether any location signal is collected; if so, keep it coarse, never treat it as evidence alone, and disclose it. Do not collect precise location.

11. **Complete a privacy review.** Map every data element to its legal basis and retention period; confirm COPPA, FERPA, SOPIPA, and state biometric-privacy treatment of keystroke-derived features; update store privacy labels and family-facing disclosures.
12. **Evaluate before scaling.** Run the study described in Section 8 and publish the results.

8. Proposed Evaluation

The claims in this report about Gold Edition's effectiveness are, at present, design hypotheses supported by indirect evidence. They should be tested. We propose four studies.

Study 1: Writing achievement. Compare students using Gold Edition with students using the previous Writing Tutor (or a delayed-start group) on on-demand writing produced under supervised conditions without AI assistance, at the beginning and end of a semester. Score essays with an analytic rubric by trained raters who do not know the students' condition. The primary outcome is change in supervised writing quality, because improvement in revised, tutor-supported work does not by itself show that students can write better unaided. A secondary outcome is the rate of spelling and grammar errors in these unassisted essays, which tests whether the conventions gate produces learning that transfers to new writing (Section 3.3).

Study 2: Revision depth. Classify changes between successive versions using the Faigley and Witte (1981) taxonomy. If the loop is working, the proportion of meaning changes should rise over the semester. Track whether bonus credit increases meaning changes or only the number of resubmissions.

Study 3: Scorer validity. Assemble a calibration set of several hundred essays across grades, scored independently by two trained teachers. Report LLM–human agreement (quadratic weighted kappa) against human–human agreement, by grade and subgroup, using established thresholds for automated scoring (Williamson et al., 2012). Run an adversarial suite: padded essays, inserted punctuation, off-topic but fluent essays, essays containing injected instructions, and essays revised by adopting the tutor's wording.

Study 4: Integrity signal accuracy. In a consented study, have volunteer students compose some essays and transcribe others from provided text, and measure how accurately process features and fingerprint signals distinguish the two, overall and by subgroup. Audit a random sample of flagged and unflagged cases with teacher review to estimate false-positive rates in practice.

Table 3. Evaluation questions, measures, and standards

Question	Measure	Standard of success
Do students write better without help?	Supervised on-demand essays before and after a semester, scored blind with an analytic rubric	Greater gain than the comparison group
Does revision become deeper?	Share of meaning changes vs. surface changes between versions (Faigley & Witte, 1981)	Share of meaning changes rises over time
Does the conventions gate teach?	Spelling and grammar error rates in new, unassisted writing	Error rates fall more than in the comparison group
Is the scorer valid?	LLM–human vs. human–human agreement, by grade and subgroup	Quadratic weighted kappa ≥ .70, and no more than .10 below human–human agreement (Williamson et al., 2012)
Is the scorer robust?	Adversarial suite: padding, punctuation, off-topic fluency, injected instructions, adopted tutor wording	No score gain from any manipulation

Question	Measure	Standard of success
Are integrity flags accurate?	Consented compose-vs.-transcribe study; audit of flagged and unflagged cases	Measured false-positive rate, overall and by subgroup, before flags reach teachers
Does bonus credit help?	Trait-level gains per round; number of rounds; student motivation survey	Bonus credit tracks improvement, not just resubmission

9. Limitations

This report describes a system in testing. Implementation details may change before release, and no outcome data from Gold Edition have yet been collected. Much of the cited research was conducted in classrooms, in higher education, or with English learners, rather than in fully online K–12 programs; findings may not transfer directly. Research on LLM tutoring and scoring is recent, much of it from 2023 to 2025, and the underlying models change faster than studies can be completed. Several cited works are preprints. Finally, the authors are the system's developers; independent evaluation is needed.

10. Conclusion

Generative AI has made the finished essay an unreliable witness to learning, but it has not changed what students need to learn: how to decide what is worth saying, organize it, and say it well. Acellus Writing Tutor Gold Edition responds by teaching and observing the process of writing rather than grading only its product. Its core instructional elements, stepwise process instruction, student-made corrections, and feedback that names the next step, rest on some of the most consistent findings in educational research. Its AI tutor follows the design that research shows separates help from harm: guide the student, never do the work. Its integrity architecture is technically strong and unusually attentive to privacy.

The work that remains is to make authorship visible in the writing itself, to keep every integrity decision in human hands, to align bonus credit with real improvement, to harden the AI scorer against the vulnerabilities that come with it, and to measure whether students become better writers when the tutor is not there. Those measurements will determine whether Gold Edition becomes what its designers intend: a model for teaching writing in the age of AI.

Acknowledgments

The author thanks the Acellus teachers, programmers, and security specialists who designed and built Writing Tutor Gold Edition, and the parents, teachers, and students whose candid feedback prompted it.

Disclosure

The author is the founder and president of the International Academy of Science, which develops and operates the Acellus Learning System and Writing Tutor Gold Edition. Gold Edition uses large language models provided by Anthropic.

References

Alexandron, G., Ruipérez-Valiente, J. A., Chen, Z., Muñoz-Merino, P. J., & Pritchard, D. E. (2017). Copying@Scale: Using harvesting accounts for collecting correct answers in a MOOC. *Computers & Education, 108*, 96–114. https://doi.org/10.1016/j.compedu.2017.01.015

Americans with Disabilities Act of 1990, Title III, 42 U.S.C. § 12181 et seq.

Apple Inc. (n.d.-a). *Establishing your app's integrity* [Developer documentation]. Retrieved November 14, 2026, from https://developer.apple.com/documentation/devicecheck/establishing-your-app-s-integrity

Apple Inc. (n.d.-b). *DeviceCheck* [Developer documentation]. Retrieved November 14, 2026, from https://developer.apple.com/documentation/devicecheck

Apple Inc. (n.d.-c). *User privacy and data use*. Retrieved November 14, 2026, from https://developer.apple.com/app-store/user-privacy-and-data-use/

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum Associates.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*(2), 201–210. https://doi.org/10.3758/BF03193441

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., Rozenberg, P., Saddiqui, S., & van Haeringen, K. (2019). Contract cheating: A survey of Australian university students. *Studies in Higher Education, 44*(11), 1837–1856. https://doi.org/10.1080/03075079.2018.1462788

California Consumer Privacy Act of 2018, Cal. Civ. Code § 1798.140 (as amended by the California Privacy Rights Act of 2020).

Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT's behavior changing over time? *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.5317da47

Children's Online Privacy Protection Rule, 16 C.F.R. pt. 312 (2025).

Coghlan, S., Miller, T., & Paterson, J. (2021). Good proctor or "Big Brother"? Ethics of online exam supervision technologies. *Philosophy & Technology, 34*(4), 1581–1606. https://doi.org/10.1007/s13347-021-00476-1

Conijn, R., Roeser, J., & van Zaanen, M. (2019). Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing, 32*(9), 2353–2374. https://doi.org/10.1007/s11145-019-09953-8

Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E., & Thies, W. (2015). Deterring cheating in online environments. *ACM Transactions on Computer-Human Interaction, 22*(6), Article 28. https://doi.org/10.1145/2810239

Crossley, S., Holmes, L., Tian, Y., Morris, W., & Choi, J. S. (2024). Plagiarism detection using keystroke logs. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 476–483). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729864

Deane, P., Roth, A., Litz, A., Goswami, V., Steck, F., Lewis, M., & Richter, T. (2018). *Behavioral differences between retyping, drafting, and editing: A writing process analysis* (Research Memorandum No. RM-18-06). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-18-06.pdf

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668. https://doi.org/10.1037/0033-2909.125.6.627

Department of Justice. (2024). Nondiscrimination on the basis of disability; Accessibility of web information and services of state and local government entities. *Federal Register, 89*, 31320–31396. https://www.govinfo.gov/content/pkg/FR-2024-04-24/html/2024-07758.htm

Department of Justice. (2026). [Interim final rule extending compliance dates for 28 C.F.R. pt. 35, subpt. H]. *Federal Register, 91*, 20902. https://www.govinfo.gov/content/pkg/FR-2026-04-20/html/2026-07663.htm

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*, Article 57. https://doi.org/10.1186/s41239-023-00425-2

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication, 32*(4), 400–414. https://doi.org/10.2307/356602

Family Educational Rights and Privacy, 34 C.F.R. pt. 99 (2025).

Federal Trade Commission. (2025). Children's Online Privacy Protection Rule. *Federal Register, 90*(76), 16918–16983. https://www.federalregister.gov/documents/2025/04/22/2025-05904/childrens-online-privacy-protection-rule

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*(4), 481–506. https://doi.org/10.3102/00346543057004481

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*, Article 1162454. https://doi.org/10.3389/frai.2023.1162454

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.58680/ccc198115885

Galletta, D. F., Durcikova, A., Everard, A., & Jones, B. M. (2005). Does spell-checking software need a warning label? *Communications of the ACM, 48*(7), 82–86. https://doi.org/10.1145/1070838.1070841

Google. (2026). *Integrity verdicts* [Play Integrity API developer documentation]. Retrieved November 14, 2026, from https://developer.android.com/google/play/integrity/verdicts

Graham, S., Bollinger, A., Booth Olson, C., D'Aoust, C., MacArthur, C., McCutchen, D., & Olinghouse, N. (2012). *Teaching elementary school students to be effective writers: A practice guide* (NCEE 2012-4058). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/practiceguide/writing_pg_062612.pdf

Graham, S., Bruch, J., Fitzgerald, J., Friedrich, L., Furgeson, J., Greene, K., Kim, J., Lyskawa, J., Olson, C. B., & Smither Wulsin, C. (2016). *Teaching secondary students to write effectively* (NCEE 2017-4002). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/508_WWCPG_SecondaryWriting_122719.pdf

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*(4), 523–547. https://doi.org/10.1086/681947

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*(4), 879–896. https://doi.org/10.1037/a0029185

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.

Graham, S., & Santangelo, T. (2014). Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review. *Reading and Writing, 27*(9), 1703–1743. https://doi.org/10.1007/s11145-014-9517-0

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security* (pp. 79–90). ACM. https://doi.org/10.1145/3605764.3623985

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hill, N. E., & Tyson, D. F. (2009). Parental involvement in middle school: A meta-analytic assessment of the strategies that promote achievement. *Developmental Psychology, 45*(3), 740–763. https://doi.org/10.1037/a0015362

Jiang, Y., Zhang, M., Hao, J., Deane, P., & Li, C. (2024). Using keystroke behavior patterns to detect nonauthentic texts in writing assessments: Evaluating the fairness of predictive models. *Journal of Educational Measurement, 61*(4), 571–594. https://doi.org/10.1111/jedm.12416

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research, 1*(1), 1–26. https://doi.org/10.17239/jowr-2008.01.01.1

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Killourhy, K. S., & Maxion, R. A. (2009). Comparing anomaly-detection algorithms for keystroke dynamics. In *Proceedings of the 2009 IEEE/IFIP International Conference on Dependable Systems and Networks* (pp. 125–134). IEEE. https://doi.org/10.1109/DSN.2009.5270346

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2506.08872

Koster, M., Tribushinina, E., de Jong, P. F., & van den Bergh, H. (2015). Teaching children to write: A meta-analysis of writing intervention research. *Journal of Writing Research, 7*(2), 249–274. https://doi.org/10.17239/jowr-2015.07.02.2

Kulik, C.-L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2), 265–299. https://doi.org/10.3102/00346543060002265

Laird, E., Dwyer, M., & Grant-Chapman, H. (2023). *Off task: EdTech threats to student privacy and equity in the age of AI*. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2023/09/091923-CDT-Off-Task-web.pdf

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Li, H., Filippov, F., Lin, Y., He, P., Yang, K., Chu, Y., Liu, H., Tang, J., & Cui, Y. (2026). *"Important" You should give me full credits!: Exploring prompt injection attacks on LLM-based automatic grading systems* [Preprint]. arXiv. https://arxiv.org/abs/2606.03090

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779

MacArthur, C. A., Graham, S., Haynes, J. B., & DeLaPaz, S. (1996). Spelling checkers and students with learning disabilities: Performance comparisons and impact on spelling. *The Journal of Special Education, 30*(1), 35–57. https://doi.org/10.1177/002246699603000103

McClain, C., Anderson, M., Sidoti, O., & Bishop, W. (2026, February 24). *How teens use and view AI*. Pew Research Center. https://www.pewresearch.org/internet/2026/02/24/how-teens-use-and-view-ai/

McKinney-Vento Homeless Assistance Act, 42 U.S.C. § 11434a.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M. H., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation and positive emotions. *Computers and Education: Artificial Intelligence, 6*, Article 100199. https://doi.org/10.1016/j.caeai.2023.100199

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75*(1), 33–52. https://doi.org/10.1037/0022-3514.75.1.33

Newton, P. M. (2018). How common is commercial contract cheating in higher education and is it increasing? A systematic review. *Frontiers in Education, 3*, Article 67. https://doi.org/10.3389/feduc.2018.00067

Ngo, T. T. N., Chen, H. H. J., & Lai, K. K. W. (2024). The effectiveness of automated writing evaluation in EFL/ESL writing: A three-level meta-analysis. *Interactive Learning Environments, 32*(2), 727–744. https://doi.org/10.1080/10494820.2022.2096642

Northcutt, C. G., Ho, A. D., & Chuang, I. L. (2016). Detecting and preventing "multiple-account" cheating in massive open online courses. *Computers & Education, 100*, 71–80. https://doi.org/10.1016/j.compedu.2016.04.008

Ogletree v. Cleveland State University, 647 F. Supp. 3d 602 (N.D. Ohio 2022).

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, Article 100234. https://doi.org/10.1016/j.caeai.2024.100234

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly, 19*(2), 139–158. https://doi.org/10.1080/10573560308222

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37* (pp. 68772–68802). https://doi.org/10.52202/079017-2197

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). Parent involvement in homework: A research synthesis. *Review of Educational Research, 78*(4), 1039–1101. https://doi.org/10.3102/0034654308325185

Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing, 21*, 104–111. https://doi.org/10.1016/j.asw.2014.05.001

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., & Gueye, B. (2011). IP geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review, 41*(2), 53–56. https://doi.org/10.1145/1971162.1971171

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior, 18*(2), 103–134. https://doi.org/10.1016/S0747-5632(01)00052-8

Rehabilitation Act of 1973, § 504, 29 U.S.C. § 794.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review, 32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. ACM. https://arxiv.org/abs/2411.16337

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sidoti, O., Park, E., & Gottfried, J. (2025, January 15). *About a quarter of U.S. teens have used ChatGPT for schoolwork – double the share in 2023*. Pew Research Center. https://www.pewresearch.org/short-reads/2025/01/15/about-a-quarter-of-us-teens-have-used-chatgpt-for-schoolwork-double-the-share-in-2023/

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Sotiriadou, P., Logan, D., Daly, A., & Guest, R. (2020). The role of authentic assessment to preserve academic integrity and promote skill development and employability. *Studies in Higher Education, 45*(11), 2132–2148. https://doi.org/10.1080/03075079.2019.1582015

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Student Online Personal Information Protection Act, Cal. Bus. & Prof. Code § 22584.

Truscott, J., & Hsu, A. Y.-P. (2008). Error correction, revision, and learning. *Journal of Second Language Writing, 17*(4), 292–305. https://doi.org/10.1016/j.jslw.2008.05.003

van Beuningen, C. G., de Jong, N. H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning, 62*(1), 1–41. https://doi.org/10.1111/j.1467-9922.2011.00674.x

Wanjura, B., Shapiro, D., Mollick, E., Mollick, L., & Meincke, L. (2026). *This is an excellent paper: The effects of prompt injection on grading* (Prompting Science Report 5). Wharton Generative AI Labs, University of Pennsylvania. https://gail.wharton.upenn.edu/research-and-insights/hidden-prompt-injections/

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity, 19*, Article 26. https://doi.org/10.1007/s40979-023-00146-z

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research, 58*(1), 87–125. https://doi.org/10.1177/0735633119830764

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features* (Research Report No. RR-15-27). Educational Testing Service. https://doi.org/10.1002/ets2.12075

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36* (Datasets and Benchmarks Track, pp. 46595–46623). https://doi.org/10.52202/075280-2020