When AI Tutors Speak: Evidence from a Randomized Field Experiment

Shihao Yang
Ph.D. Candidate Boston University
syang536@bu.edu

Marshall Van Alstyne
Questrom Professor of Information Economics
mva@bu.edu

Chis Dellarocas
Richard Shipley Professor of Digital Learning & Innovation
dell@bu.edu

Abstract

Expert-configured AI tutors can extend structured practice and coach based on performance, but their contribution to learning outcomes and the value of spoken interaction require field evidence. We report a five-week randomized experiment with 86 working professionals in an asynchronous online MBA finance module. Students were offered a course-specific tutor (n = 52) or assigned to a holdout group (n = 34). Tutor students were further randomized to voice and text modes that alternated across weeks, holding constant the knowledge base and visual blackboard. Among 46 students in the retained assessment sample, tutor access was associated with higher post-test scores (+3.6 to +6.6 of 55 points depending on covariates) and greater written-answer depth. The instructor's final showed a +2.6-percentage-point tutor access boost across the full cohort. Voice increased dialogue density, question frequency, and message length, with similar engagement time. No voice assessment advantage was detected and the five-week sensitivity estimates were near zero. Observed voice billing was 2.8 times text billing. Evidence suggests that structured AI tutoring significantly improves learning outcomes, and students found it a valuable added resource. Voice, however, expanded participation without a detected assessment benefit relative to text.

Keywords: Intelligent Tutoring Systems, Adult Learning, Distance Education, Online Learning, Human-computer Interaction, AI, Evaluation Methodologies

1. Introduction

How much can tailored AI tutoring raise academic performance compared to traditional study or to use of commercial AI? New generative AI tutors can extend student practice, vet their explanations, sequence hints, and interactively test their applied knowledge. Their educational value may also depend on students' modes of interaction, as speaking and writing might shape their participation differently. This research reports on a field experiment that explores two lines of inquiry (i) whether a course specific generative-AI tutor improves learning outcomes compared to existing resources, and (ii) whether text versus voice delivery modes affect interaction levels, preferences, performance, and delivery cost.

Generative-AI experiments demonstrate the importance of instructional configuration. Bastani et al. (2025) find that assistance improves practice performance but can impair subsequent unaided performance. Learning safeguards can mitigate this harm. Kestin et al. (2025) report greater immediate learning with a configured tutor than active classroom instruction. These different interventions direct attention to how assistance organizes retrieval, explanation, and application.

Reporting on the University of Maryland's evaluation of its Virtual Study Assistant, Palmer (2026) describes low student uptake, instructors' limited use of tutor mode, and access to competing AI tools.

These conditions make adoption and instructional implementation central to evaluating campus AI deployment.

Field experiments on generative AI in work and education settings examine temporary upskilling (Wiles et al., 2026), analytics support (Bapna et al., 2025), and personalization (Agrawal et al., 2026); graduate tutoring raises the corresponding question of what students can demonstrate beyond assisted practice.

Working professionals study around employment and family responsibilities. Our tutor supplemented an asynchronous course's existing resources. Online-education research emphasizes sustained participation (Reich & Ruipérez-Valiente, 2019) and associates active practice with stronger performance than passive viewing (Koedinger et al., 2015).

Reviews find that tutoring systems support learning, with effects depending on comparison instruction, assessment, and implementation (Kulik & Fletcher, 2016; Ma et al., 2014; VanLehn, 2011). New systems therefore require fidelity evaluation and locally validated outcomes. Screened ChatGPT-authored mathematics hints also yield learning gains (Pardos & Bhandari, 2024), through curated help rather than unrestricted conversational tutoring.

Earlier dialogue-tutor experiments provide a modality benchmark. Litman et al. (2006) observe more spoken turns without a reliable adjusted assessment advantage. D'Mello et al. (2011) find no overall speech advantage and propose that typing facilitates formulation. Modern models permit evaluating flexible dialogue, spoken output, and repeated course-grounded use together.

We conducted a five-week randomized field experiment with 86 online MBA professionals studying pricing, customer analytics, and corporate finance. The allocation used pre-test-ranked blocks. Tutor students were randomized to a starting modality and alternated voice and text. Both modes shared the model, knowledge base, transcript, and persistent blackboard. Several key results emerged.

First, students who used the generative AI tutor had higher post-test scores (+3.6 to +6.6 of 55 points, depending on adjustment). Their written answers also exhibited higher relationship quality among ideas (interconnected answers 49% versus 27%). An independent test, the instructor’s final, showed a +2.6-percentage-point access contrast across all 86 students.

Second, this study extends the comparisons of Litman et al. (2006) and D'Mello et al. (2011) to a generative tutor with fluent spoken output, repeated course use, and shared visual support. In our context, voice increased turn density for 31 of 32 students, with more questions and longer messages at similar engaged time. No weekly assessment advantage was detected: the prioritized Weeks 8–9 contrast was −.08 of 11 points (95% CI [−1.10, .94]). Subsequent weeks showed no statistical difference in performance.

Third, the field test separately measures participation, assessed performance, student preferences, and delivery cost. Voice roughly doubled dialogue density; student preferences shifted toward voice over text; voice billing was 2.8 times text billing at observed use; yet no assessment difference was detected. Institutions can therefore weigh participation and user experience against expenditure when assessing net benefits.

Fourth, this study documents attempt-before-explanation, single-question sequencing, graduated help, structured feedback, and transfer-capstone requirements for subsequent fidelity and mechanism studies.

Taken together, the findings are consistent with educational benefits from the structured, incentivized practice bundle the tutor made available, with no detected additional assessment benefit from voice versus text mode.

The remaining paper is organized as follows: Section 2 builds from prior research to organize hypotheses. Section 3 presents research design and metrics. Section 4 covers findings and robustness. Sections 5 and 6 interpret findings and discuss implications. Section 7 covers limitations, and Section 8 concludes.

2. Prior Research and Hypotheses

2.1. Pedagogical orchestration and the allocation of learning work

Tutoring allocates diagnosis, attempts, hints, feedback, and application around learners' activity (Koedinger & Aleven, 2007; VanLehn et al., 2003). Premature explanations remove reasoning opportunities; insufficient help prevents progress. A course-grounded workflow sequences attempts and graduated support to sustain productive work.

Retrieval practice supports retention (Karpicke & Blunt, 2011; McDermott, 2021; Roediger & Karpicke, 2006), and self-explanation improves understanding (Chi, 1996; Chi et al., 1994). The interactive, constructive, active, passive (ICAP) framework distinguishes whether students passively receive information or manipulate and engage with it (Chi, 2009; Chi & Wylie, 2014). The framework also measures cognitive activity level, moving from recording to synthesizing and to applying beyond what was presented. Output measures participation, explanations connect principles, and inferences provide evidence of constructive activity.

Self-regulated learning entails judging understanding, selecting practice, monitoring progress, and seeking help. Tutor questions and explanation checks can scaffold these activities (Azevedo, 2005; Azevedo & Hadwin, 2005).

Student-agency and learner-control research (Darvishi et al., 2024; Qin et al., 2025) and simulated practice (Mollick et al., 2024) motivate specifying learners' decisions and revision opportunities during AI-supported tasks.

Cognitive offloading reallocates work to external support (Risko & Gilbert, 2016). Tutors can also reduce organizational effort while preserving reasoning or remove practice by supplying answers. Default answers may not adapt to student needs (LearnLM Team, 2024). Misuse of technology risks harming knowledge encoding and long term retention.

Internet access changes memory strategies and can inflate perceived internal knowledge (Fisher et al., 2015; Sparrow et al., 2011). Technology can alter cognitive effort (Lee et al., 2025), and offload critical-thinking (Gerlich, 2025). Others have studied electrophysiological essay-writing (Kosmyna et al., 2025), and longitudinal secondary-school learning penalties (Stromberg et al., 2026). Such studies motivate assessing capability beyond assisted practice and seek to facilitate the reasoning learners perform.

Bloom presents one of the more established taxonomies of learning objectives based on how learners acquire new knowledge in cognitive (knowledge), affective (emotional), and psychomotor (action) domains (Bloom 1956). Anderson and Krathwohl expanded the framework to action ranges such as recalling, explaining, analyzing, and creating (Anderson & Krathwohl, 2001). The Structure of Observed Learning Outcomes (SOLO) framework complements this identifying whether student understanding is prestructural (disconnected or missing the point), unistructural (grasping a single relevant aspect),

multistructural (knowing multiple but independent points), relational (connecting different points), and extended (generalizing principles to new domains (Biggs & Collis, 1982). Our field research ranks higher and lower level item understanding using standard Bloom tags, while we use SOLO scores to characterize written assessment quality.

2.2. Spoken and typed tutoring

As distinct forms of interaction, speech and writing can differ in knowledge density and exchange. Litman et al. (2006) find more turns and longer sessions with speech, but no reliable adjusted assessment or words-per-turn difference. Recognition errors also complicate interaction. D'Mello et al. (2011) use human transcription in counterbalanced experiments and find no overall modality advantage, but do find a typing advantage among highly motivated learners. Their tutor spoke in both input conditions, underscoring the need to specify input and output modality.

Multimedia modality benefits vary with pacing and element interactivity (Ginns, 2005). Spoken explanations can integrate words and diagrams, while duplicate displayed text can reduce the benefit (Kalyuga et al., 1999; Mayer et al., 2001). Transcripts and blackboards support review in both modes. A text transcript duplicates speech and may add processing demands.

Lachner et al. (2018) find that written explanation supports organization and conceptual knowledge, while oral explaining supports elaboration and transfer in a non-tutoring task. Video explanation benefits learning relative to restudy but includes cues beyond speech (Hoogerheide et al., 2016). Explanation organization, elaboration, and accuracy therefore require measures distinct from volume.

Writing facilitates planning before production; typing permits formulation before submission. By contrast, speech can reduce production effort and invite spontaneous questions. Answer production and explanation comprehension impose different demands. Shared transcripts and blackboards permit review during either spoken or typed exchange.

Speaking studied material can improve memory relative to silent reading, through recollection and familiarity (MacLeod & Bodner, 2017; MacLeod et al., 2010; Ozubko et al., 2012). This production-effect account directs attention to the explanations, retrieval attempts, and justifications students articulate during tutoring.

Jacobs and Kiesler (2025) examine voice mode tutors for novice programmers with disabilities. Students found it competent but it had limited correctness uttering code elements. Riegel et al. (2026) studied voice input for programming but found most students used and prefered typing. Our configuration combines spoken input/output, shared written support, and repeated finance practice, defining a different modality comparison.

2.3. Competing accounts of modality and the predictions they imply

Four accounts make different predictions for spoken versus typed tutoring. Table 1 summarizes their mechanisms and observable implications.

Table 1: Four accounts of modality effects in tutoring and their predictions

Account	Mechanism	Prediction for voice relative to text	Observable in this study

Account	Mechanism	Prediction for voice relative to text	Observable in this study
Social agency	A human-like voice is a social cue that raises engagement and effort to understand (Atkinson et al., 2005; Mayer et al., 2003; Moreno & Mayer, 2000; Nass & Brave, 2005; Reeves & Nass, 1996)	More participation and higher assessed learning	Turn density, questions, preference; weekly subscores
Production effect	Articulating material aloud strengthens later memory for it relative to silent processing (MacLeod et al., 2010; Ozubko et al., 2012; MacLeod & Bodner, 2017)	Higher assessed learning, concentrated in content the student voiced	Weekly subscores; words per message
Text facilitation	Typing allows planning before submission; speech favors immediate, less formulated responses (D'Mello et al., 2011; Lachner et al., 2018)	More turns in voice but equal or lower assessed learning; longer reply gaps in text	Reply gaps, pre-keystroke time; weekly subscores
Shared visual channel	Both modes provide a live transcript and a persistent blackboard; the modality effect in multimedia learning depends on pacing and on redundant text (Ginns, 2005; Kalyuga et al., 1999; Mayer et al., 2001)	Any modality difference is attenuated	Weekly subscores; depth scores

Social agency and production effort predict a voice learning advantage; text facilitation predicts greater participation without that advantage; shared visual support predicts attenuation. These accounts differ most clearly on the within-student assessment contrast.

Social-actor and media research link speaking systems to interpersonal expectations (Nass & Moon, 2000; Nass et al., 1994); anthropomorphism identifies triggering cues (Epley et al., 2007); voice experiments show increased perceived intellect and humanness (Schroeder & Epley, 2015, 2016; Seaborn et al., 2021). Applied to tutoring, spoken exchange should sustain participation and thereby learning. A voice assessment advantage would favor social-agency or production accounts; a participation difference without a detected advantage would be consistent with text facilitation or shared-channel attenuation, resembling earlier dialogue-tutor findings. Because these accounts share observable signatures, turn density, reply gaps, pre-keystroke time, and message length accompany the assessment contrast. They describe response patterns, not cognitive activity directly (Chi & Wylie, 2014).

Experience differs from learning. Voice-assistant and conversational-interface studies measure usability, engagement, adoption, and reuse (Chen et al., 2024; Fernandes & Oliveira, 2021; McLean & Osei-Frimpong, 2019; Moriuchi, 2019, 2021; Qiu & Benbasat, 2009; Rzepka et al., 2022; Seeger et al., 2021; Wagner et al., 2019); children's agent studies examine contingent dialogue, engagement, and comprehension (Xu et al., 2022a, 2022b). These constructs explain preference and continued use without establishing superior learning. Our study adds to this literature by measuring preferences, participation, and performance separately.

2.4. Research Hypotheses

Our primary research question (RQ1) asks how offering a class-specific AI tutor prompted with learning pedagogy affects student learning alongside existing instruction and resources. Our secondary research question (RQ2) asks how voice and text differ in observed interaction, topic-specific performance, student preference, and delivery cost. We expand these into the following hypotheses:

- H1: AI tutor access increases total assessed performance relative to the holdout, controlling for baseline performance.
- H2: Within the tutor arm, voice produces higher performance than text on assessment items associated with the relevant course week.
- H3: The access effect is larger for higher-order than lower-order Bloom and SOLO levels.
- H4: Learning gains increase with tutor engagement intensity.

The first hypothesis examines whether the randomized treatment group experiences gains relative to the holdout group. The second hypothesis extends analysis to randomization across voice and text modes. The third hypothesis predicts greater gains for more sophisticated understanding as distinct from mere recall. The fourth hypothesis explores whether gains accumulate.

3. Research Design and Measurement

3.1. Setting and participation

The field experiment ran during Weeks 8–12 of an online MBA module covering pricing, customer analytics, and corporate finance. Generative AI tutoring was available from July 1 to August 5, 2026, and the post-test window closed on August 15. We trained the tutor on course specific content including recorded lectures, readings, problem sets, discussion boards, and instructor exchanges. Participation was voluntary and carried no course credit. Holdout students retained standard course resources and access to consumer AI. The access contrast therefore evaluates the offer of an additional configured tutoring resource within the existing course environment.

The study received institutional review board approval on June 19, 2026 (institution and protocol number withheld for anonymous review). Participants gave informed consent before randomization.

Incentives were identical across arms: $4 per completed study week (tutoring practice in the access arm; a brief check-in survey in holdout), rising to a $20 completion payment at four completed weeks plus a $5 bonus for completing all five weeks. Payments earned per week were retained regardless of post-test completion. Completing the post-test entered students into a raffle of $600, $400, and $200 prizes and earned a free e-book authored by program faculty. Participation measures describe use under these incentives.

The cohort comprised 86 assigned students: 52 offered tutor access and 34 assigned to holdout. The recorded response-quality screen retained 46 assessments (29 tutor; 17 holdout), excluding four students for short pre-test or post-test duration and one for short-answer effort. Table 2 separates assignment, completion, and analysis denominators.

Table 2: Study population and analysis samples

Population or record	Count	Interpretation
Assigned participants	86	52 tutor access; 34 holdout
Initial tutor-mode randomization	51	25 voice-start; 26 text-start (one student lacked a recorded start mode)
Paired pre/post completers	51	32 tutor; 19 holdout (coincidentally also 51; a different set from the row above)

Population or record	Count	Interpretation
Screened assessment sample	46	29 tutor; 17 holdout
Five-week H2 panel	145 student-weeks	29 students × 5; includes unattended weeks
Behavioral contrasts	32 paired students	Students observed in both modes for these measures
Language contrasts	34 paired students	Different available-case sample from the assessment panel
Archive-based paired density contrasts	27 students	Archive coverage and pairing determine inclusion
Weekly tutor-arm surveys	97 responses; 32 students	27 repeat responders
Active tutor users	38	At least one session
Field-period conversation archive	256 conversations	298 application session rows are a different unit

3.2. Allocation and the modality contrast

The allocation used pre-test-ranked five-student blocks, with three tutor and two holdout assignments in complete blocks. The reconstructed analysis blocks follow the historical scoring convention; assignment documentation is incomplete. Recorded starting-mode allocation covers 51 tutor students (25 voice-start; 26 text-start). The modality analysis uses 29 students with retained post-tests and mode histories.

In Weeks 8 and 9, students followed their randomized starting modality and calendar alternation. From Week 10, a missed week's mode carried forward to the next attempt. Five-week modality histories therefore combine the randomized starting sequence with subsequent completion. In the present analysis, we prioritize the directly randomized Weeks 8–9 contrast for interpreting H2 and treat the five-week panels as sensitivity analyses. We code the original calendar schedule, platform-assigned mode, served mode, attendance, and goal completion separately. Student fixed effects account for stable learner differences; course-week effects account for common topic differences.

The modality estimand concerns weekly voice–text differences within an alternating, mixed-mode regime. Students could review or preview content across weeks, so some carryover between topics is possible.

The screened modality panel contains 13 voice-start and 16 text-start students. Analysis coding matches platform assignment and all attended exposures. Calendar assignment differs in seven later weeks across five students (Online Appendix Section C).

Figure 1 summarizes participant flow in CONSORT form, and Figure 2 displays observed pre-test distributions.

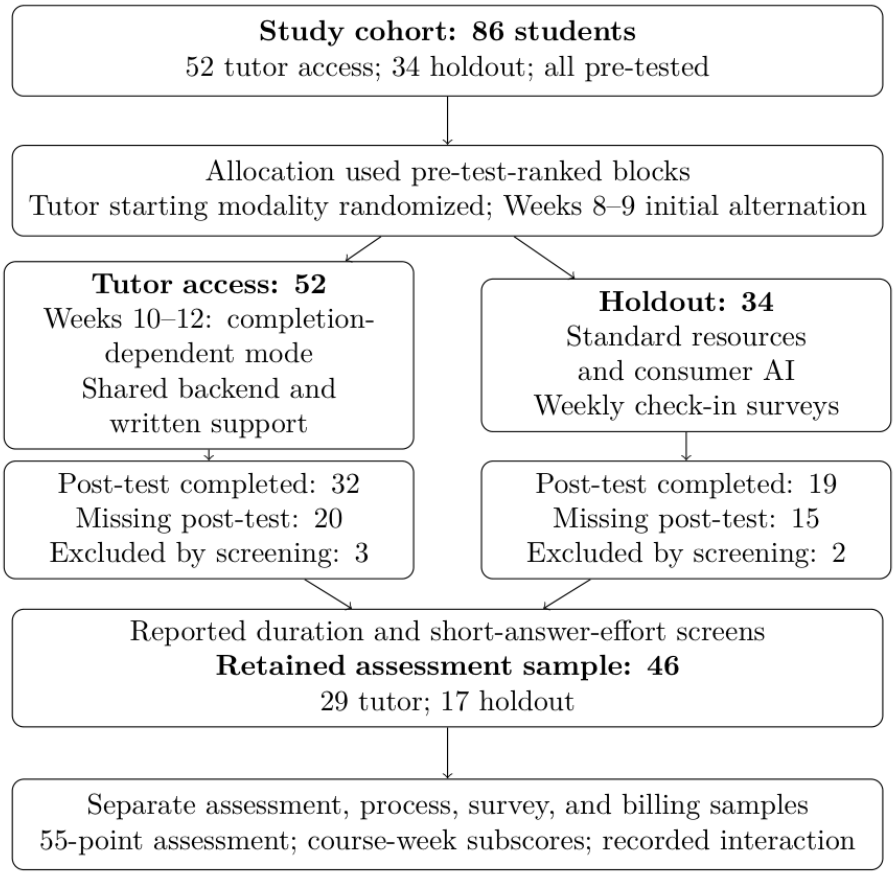


Figure 1: Study design and participant flow for the 86-student cohort and completion-dependent modality regime. Study enrollment and retained analysis denominators are shown separately. Screening removed four duration failures and one effort failure overall. Prepared in TikZ with coding assistance from Codex (OpenAI; GPT-6.1-sol).

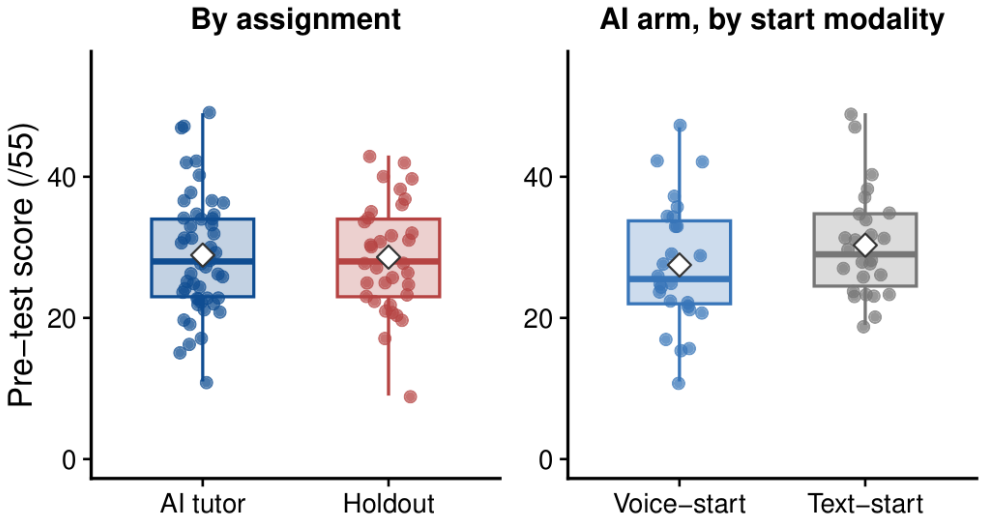


Figure 2: Observed pre-test score distributions in the 86-student study cohort, grouped by tutor access and recorded starting modality. Boxes show medians and interquartile ranges; diamonds show means.

3.3. System configuration and intended pedagogy

The protocol specified course-grounded practice, attempts before explanations, one question at a time, adaptive difficulty, graduated help, five-part feedback, and a scaffolded transfer capstone. Table 3 distinguishes intended requirements from implementation evidence; fidelity and accuracy were not independently audited. Figure 3 illustrates the sequence.

Table 3: Intended tutoring design and available evidence

Design element	**Intended instruction**	**Learning rationale / hypothesis**	**Model Behavior**
Attempt before explanation	Request an attempt before help	Retrieval and explanation; H1	Prompted; fidelity unaudited
Question sequence	One question; adapt demand	Sustain student production	Prompted; fidelity unaudited
Support when stuck	Hints or smaller steps; example then practice	Balance help with activity	Prompted; fidelity unaudited
Feedback	Verdict, strengths, improvement, structure, micro-hint	Support revision	Sequence described; unaudited
Transfer capstone	Generalize with scaffolding	Integration and transfer	Prompted; fidelity unaudited
Modality	Speech or text with blackboard	Response production; H2	Mode logged; cognitive activity unvalidated
Nudge and weekly practice	Re-engage; meter weekly goal	Participation; H4	Prompted; events unaudited

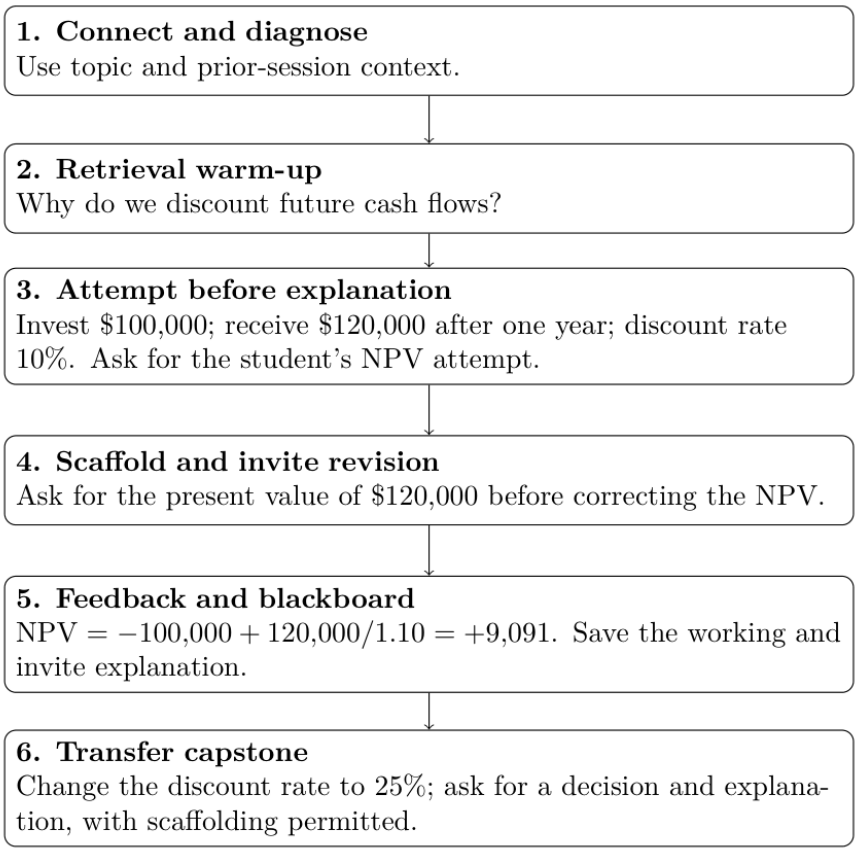


Figure 3: Intended tutoring cycle illustrated with an NPV practice problem. The sequence specifies the intended instructional design. Prepared in TikZ with coding assistance from Codex (OpenAI; GPT-6.1-sol).

Both modes used GPT-5.4 and the same course knowledge base, transcript, and persistent blackboard. Voice transcribed speech, passed text to the model, and synthesized replies with live captions; the model did not receive audio, intonation, or affect. Text thinking time and compose telemetry began in Weeks 11–12. Online Appendix Figure S4 documents the architecture; platform and course identifiers are omitted for anonymous review.

Figure 4 shows the deployed session interface. Online Appendix Figures S2 and S3 show the student and instructor dashboards.

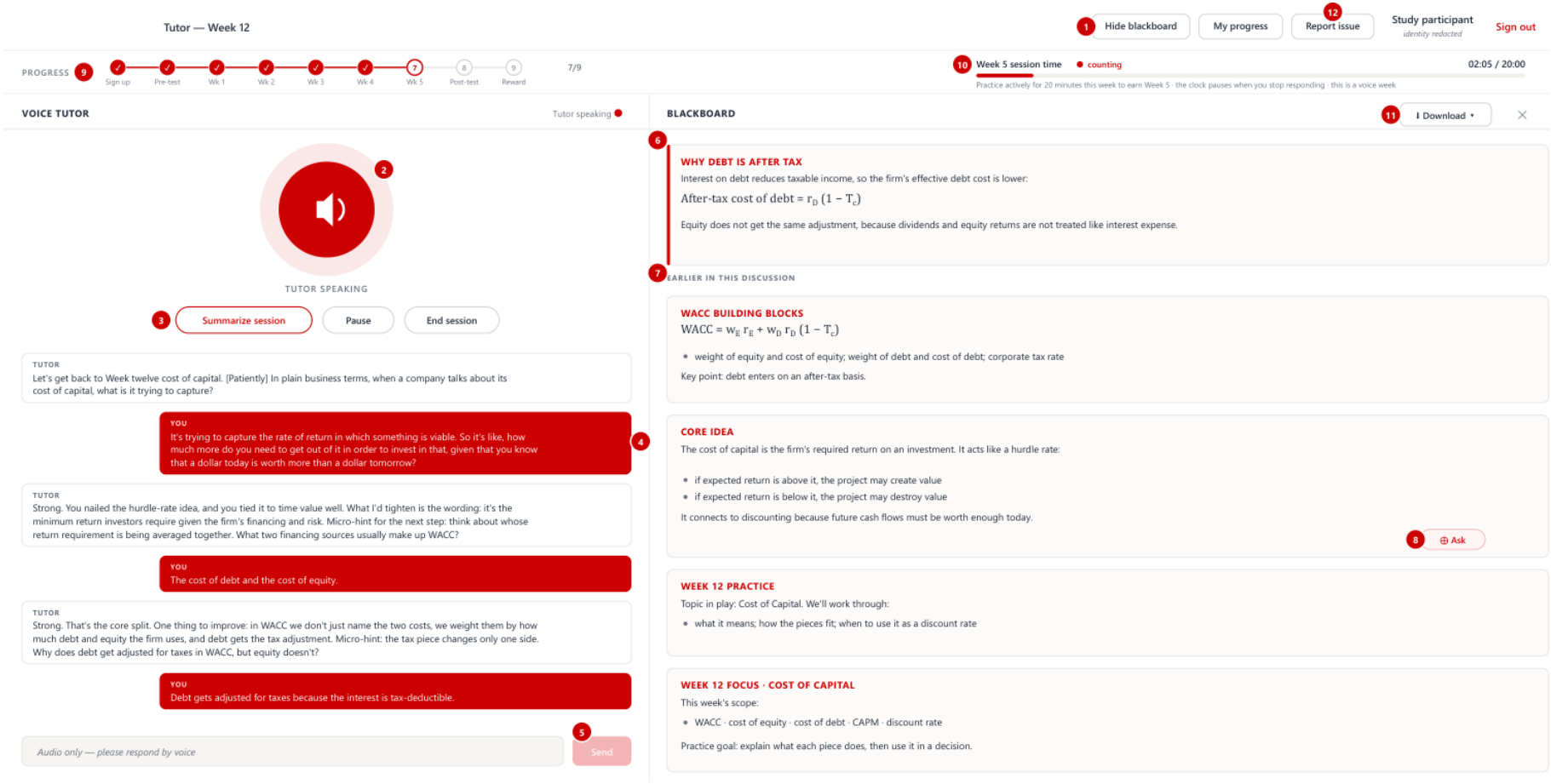


Figure 4: Anonymous voice-tutor interface, shown in full. Callouts: 1 blackboard visibility; 2 speech status; 3 session controls; 4 transcript; 5 input control; 6 current blackboard note; 7 earlier notes; 8 ask-about-note control; 9 study progress; 10 weekly time counter; 11 note download; 12 issue reporting. Identity and branding removed.

The access contrast evaluates the configured bundle as a whole.

3.4. Assessments and scoring

Each study assessment contained 15 items: two multiple-choice questions and one short answer for each of five course weeks. Multiple-choice items contributed up to 30 points and short answers up to 25, yielding a total score out of 55. Each course-week subscore ranged from 0 to 11 and represented the corresponding content on the end-of-study assessment.

Human reviewers checked the construct alignment, wording, distractors, and answer keys of LLM-assisted items. Paired variants changed surface details while preserving the target construct; post-tests used complementary vectors. Online Appendix Section B describes the variant construction and its verification limits.

Bloom tags describe a question's cognitive demand (Anderson & Krathwohl, 2001); SOLO scores describe the structure of the student's written response (Biggs & Collis, 1982). SOLO ranges from prestructural (0) to extended abstract (4); the student dashboard displayed the same levels on a 1–5 scale. The two depth endpoints are the mean SOLO level and the proportion of answers reaching relational quality or above (SOLO ≥3). For the total-score calculation, SOLO levels map to 0, 1, 2.5, 4, and 5 points. Distinguishing item demand from response quality permits separate examination of cognitive-demand contrasts and written-answer depth.

Table 4 summarizes the two classifications. Online Appendix Figure S1 illustrates complementary variant vectors.

Table 4: Taxonomies used for item demand and written-answer quality

Dimension	Bloom item demand	SOLO answer quality (level)
Lower	Remember; understand; apply	Prestructural (0); unistructural (1); multistructural (2)
Higher	Analyze; evaluate; create	Relational (3); extended abstract (4)

Note. The columns are distinct classifications, not a conversion between Bloom and SOLO levels. SOLO grades are mapped to points as described in the text.

Automated-scoring research distinguishes reliability from validity. Machine and human scoring can differ in severity and sensitivity to task context (Zhai et al., 2020), and LLM scoring depends on model, examples, rubric, and item context (Lee et al., 2024). These dependencies matter especially where tutoring feedback may influence answer length or structure.

Short answers were scored by a masked council of four commercial language models, each applying the same SOLO rubric; the lower median of the four scores was taken to guard against a single generous grader. Mean linearly weighted agreement among the four graders was $\kappa = .81$, with 16% of answers showing disagreements spanning at least two SOLO levels. Online Appendix Section D reports agreement and provider-sensitivity summaries.

3.5. Instructor-course midterm and final instruments

Both instructor instruments are observed for all 86 students. Raw midterm scores out of 150 and final scores out of 200 were normalized to 0–100; contrasts are percentage points. Midterm administration occurred July 3–11 during the intervention but covered material pre intervention. Final exam

administration occurred August 7–13, before or overlapping research post-testing. The 46-student final analysis uses retained study-assessment students.

Course-total ANCOVAs adjust for study pre-test and ability blocks with HC1 errors; item contrasts use item/block effects and student-clustered errors. These course outcomes were not pre-specified and are exploratory. We report both totals and retain item-demand contrasts in Online Appendix Table S2. The full-cohort final avoids study-post-test attrition; its retained-sample counterpart bridges to the main assessment.

3.6. Process, preference, and cost measures

We measure attendance, goal completion, engaged time, turn density, questions, and reply timing with their respective exposure denominators. Engaged time applies an idle clamp and a 20-minute weekly goal. Conversation rates are defined for attended sessions.

Conversation logs record student and tutor turns, timestamps, and message text. An LLM classifier labels substantive turns, filler, substantive questions, and clarification requests; these categories have not been validated against human coding. Message length, question counts, turn density, and reply gaps measure observable conversational activity.

Voice timing subtracts estimated tutor playback from the reply interval. Text telemetry separates pre-keystroke and composition time. These instruments describe distinct components of student response production.

Surveys elicited experience ratings and modality preference. The preference scale ranged from −2 (text only) to +2 (voice only). Billing used connected voice minutes and text messages; these units differ from engaged time and student turns.

3.7. Estimation

Access ANCOVAs regress post-test score on tutor assignment and pre-test, with and without reconstructed allocation-block indicators. Historical OLS estimates are retained; HC1 and leverage-adjusted sensitivities appear in Online Appendix Section M. A July 7 repository version specifies blocks and robust errors, but the deposited analysis plan remains unverified. Missingness analyses include IPW, Lee point bounds, and approximate multiple imputation. Raw points remain primary; adjusted Hedges' g uses the pooled within-arm post-test SD.

H2 regresses course-week subscores on voice mode with student and course-week fixed effects and student-clustered errors. The interpretation prioritizes the directly randomized Weeks 8–9 calendar assignment; three sensitivity panels use five-week platform assignment, attended served mode, and all-week calendar assignment (Online Appendix Table S1). Attendance-conditioned panels concern participants with observed exposure.

Modality inference uses finite-sample-corrected student-cluster covariance and t_{G-1} tests. We report estimates and confidence intervals without interpreting a null contrast as educational equivalence.

The main tables report raw p-values. A post hoc Holm sensitivity covers total score and the two depth endpoints; secondary analyses retain the Benjamini–Hochberg family. Course records, process measures, and surveys are exploratory or descriptive.

4. Results

4.1. Participation and exposure

Thirty-eight students used the tutor; 14 students offered the tutor recorded no session. Weekly completions and check-ins appear in Table 5. The 29-student modality panel contains 145 student-weeks, including 124 attended weeks and 118 completed weeks; two retained students never attended.

Table 5: Field-period study records

Record	Count
Study participants	86 (52 tutor; 34 holdout)
Weekly tutor completions	29 / 31 / 28 / 30 / 26 of 52
Weekly holdout check-ins	20 / 23 / 21 / 22 / 19 of 34
Completed at least four weeks	27 tutor; 20 holdout
Tutor conversations	256 (298 application session rows)
Student messages	4,231 (2,526 voice; 1,705 text)
Tutor-arm weekly surveys	97 responses from 32 students
Instrumented typed turns, Weeks 11–12	621

Note. Field period July 1–August 5, 2026; post-study usage excluded. Conversations, messages, application sessions, and billed units have distinct denominators.

Post-test completion was 32/52 in the tutor arm and 19/34 in holdout. The retained assessment sample represents 46/86 assigned students (53.5%): 35 did not complete a post-test and five were removed by screening. Retention was 29/52 in the tutor arm and 17/34 in holdout. Section 4.2 examines who was retained.

4.2. Attrition and sample composition

Retained students had higher workbook-baseline scores in both arms: 31.34 versus 25.83 in the tutor group and 31.35 versus 25.88 in the holdout group (Online Appendix Table S3). Arm retention rates did not differ detectably (Fisher $p = .662$). Within tutor, retention was 27/38 among users and 2/14 among nonusers. Table 6 records these counts.

Table 6: Retention and observed tutor use by arm

	Tutor arm: retained (n = 29)	Tutor arm: not retained (n = 23)	Holdout: retained (n = 17)	Holdout: not retained (n = 17)
Students, n	29	23	17	17
Used the tutor at least once, n	27	11	—	—
No recorded tutor session, n	2	12	—	—

Note. Retained assessments passed the recorded duration and effort screens. Tutor-use counts concern the offered arm.

The historical Inverse Probability Weighting (IPW) estimate was +6.52. Identifier-linked full-cohort weights yielded +6.54 (HC1 95% CI [2.20, 10.88]). Under monotonic retention, unadjusted Lee point bounds were [2.51, 5.53] for paired completers and [2.45, 5.18] after screening. Approximate imputation across all 86 students yielded block-outcome estimates of +3.37 (95% CI [−1.47, 8.22]) using a pre-test

reference and +5.04 [−.24, 10.32] using a block reference. These distinct sensitivities and their assumptions are detailed in Online Appendix Section M.

4.3. Recorded interaction differed substantially across modes

Observed voice sessions had greater student turn density, more questions, and longer messages, with similar engaged time per completed week (Table 7). Exact paired sign tests assess the direction of individual differences; the table displays the corresponding process summaries.

Table 7: Paired interaction measures

Measure	Voice	Text	n	Within-student direction / p
Engaged minutes per completed week	27.5	28.0	32	14 voice, 17 text, 1 tie; .72
Student turns per conversation minute	1.34	0.75	32	31/32 higher in voice; < .001
Accumulated questions per student	11.3	4.6	32	25/32 higher in voice; .002
Words per student message	23.7	16.1	34	28/34 higher in voice; < .001
Student median raw reply gap (seconds)	27	54	27	26/27 faster in voice; < .001

Note. n is paired students. Available-case coverage differs by measure. Sign-test p-values concern individual directions rather than the displayed means. Questions are accumulated counts, and raw reply gaps include different activities in the two modes.

In the separate 27-student paired process sample, voice turn density exceeded text density for every student, and classifier-labelled substantive density was higher for 26 of 27 (Figure 5). Classifier-labelled activity increased alongside total activity. Online Appendix Section F describes the samples, classifications, and aggregation.

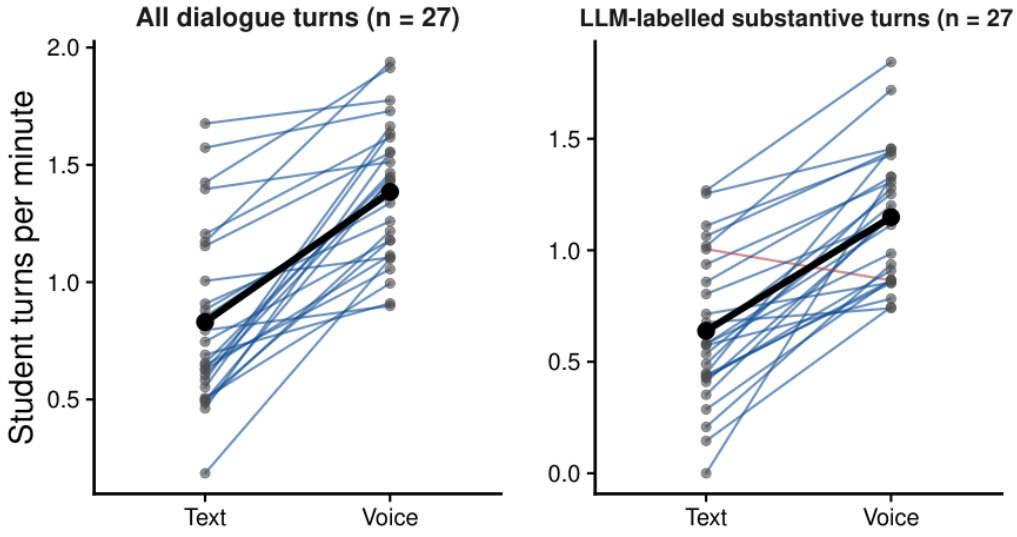


Figure 5: Student-by-student paired turn densities in the separate process sample in text and voice. Left: all dialogue turns (27 of 27 higher in voice). Right: classifier-labelled substantive turns (26 of 27 higher in voice). The archive sample differs from the 32-student behavioral contrast in Table 7.

Raw student-level median reply gaps were 27 seconds in voice and 54 in text (Figure 6). The appendix reports playback-adjusted voice gaps and text pre-keystroke/composition intervals. Negative adjusted voice gaps reflect overlap or playback-estimation error. Faster spoken replies and longer pre-keystroke pauses in text are the pattern the text-facilitation account predicts (Section 2.3).

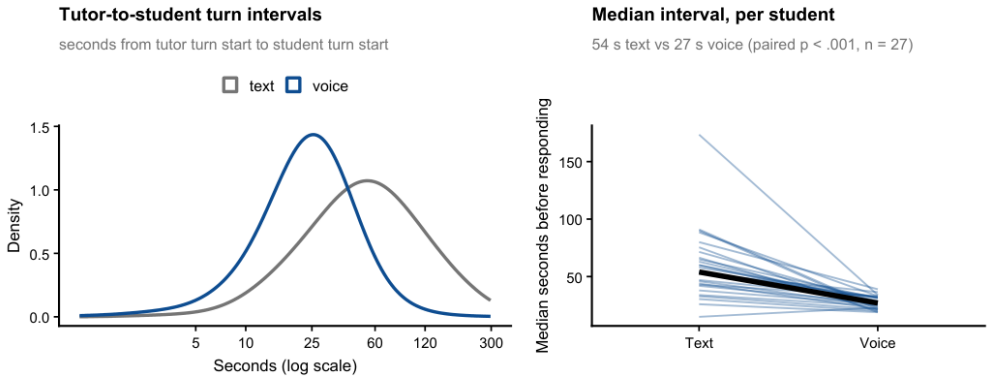


Figure 6: Tutor-to-student raw turn intervals from the paired process display. The plot rounds student medians to 27 seconds in voice and 54 seconds in text (27 paired students). These intervals combine tutor output and student reply timing, rather than directly measure thinking effort or system latency.

Completion ran in the opposite direction from turn density. In the final-week subgroup, 17 of 18 text opportunities but only 9 of 12 voice opportunities were completed under completion-dependent assignment, a pattern we return to in Section 5.4. The within-student text-pause analysis for Weeks 11–12 was not estimable: no student in the eligible panel had usable text intervals in both weeks.

4.4. Stated preference and experience

Weekly experience surveys yielded 97 deduplicated, week-labelled responses from 32/52 tutor students. Holdout coverage was 27/34 students with 121 finished records; absent week labels prevent student-week counts. Weekly tutor preference means were −.21, +.05, +.11, +.69, and +.29 showing a steady rise in willingness to use. Among 27 repeat respondents, 12 moved toward voice, 10 were unchanged, and five moved toward text.

These preference trajectories describe respondents under the alternating-mode regime. The mode served in a given week may influence that week's rating, and the process, preference, and assessment samples overlap only partially, so the combined pattern does not show that students who preferred voice learned more with it.

4.5. Assessment performance

Mean gains were 15.1 points in tutor and 11.4 in holdout. On the same 46 students and outcome, adding blocks changes the conditional comparison: pre-test-only OLS gives +3.56/55 (95% CI [−.43, 7.54], p = .079); block-adjusted OLS gives +6.63 [1.95, 11.31] (p = .007). Adjusted Hedges' g is .48 and .90, respectively, with approximate bootstrap intervals [−.001, 1.08] and [.29, 1.62]. A coefficient-preserving leverage-adjusted sensitivity gave an HC3 interval of [1.72, 11.54] for the block estimate (Online Appendix Section M).

Table 8 reports the assessment estimates, and Figure 7 displays total-score and written-depth estimates on their original scales.

Table 8: Assessment estimates

Contrast and outcome	Estimate	95% interval	Raw p
H1: access, total /55, pre-test adjustment	+3.56	[−0.4, +7.5]	.079†
H1: access, total /55, plus blocks	+6.63	[+1.95, +11.31]	.007***
Access, short-answer component /25	+2.86	[+0.14, +5.6]	.040**
Access, multiple-choice component /30	+0.88	[−1.5, +3.3]	.46
Depth: share SOLO ≥ 3	+.221	[+.02, +.42]	.029**
Depth: mean SOLO level	+.391	[+.01, +.77]	.046**
H2: voice minus text, Weeks 8–9 (prioritized), subscore /11	−.079	[−1.103, +.944]	.875
H2: voice minus text, five-week platform panel	−.006	[−.616, +.604]	.985
H3 (exploratory): access by higher-order Bloom	+.06	[-.10, +.21]	.48
H4 (exploratory): gain per completed week	+2.00	[+.37, +3.6]	.018**

Note. Historical access and depth intervals use OLS errors. Post hoc Holm p-values for block-adjusted H1 and two depth outcomes are .021, .058, and .058, using rounded OLS values. H4 concerns self-selected completion.

Retaining the effort-excluded assessment yielded +5.52 (OLS p = .052); alternative-provider scoring yielded +5.22 (p = .044). Access estimates depend on adjustment, screening, and scoring.

Pre-test-adjusted component estimates were +2.86/25 for short answers (95% CI [.14, 5.58], p = .040) and +.88/30 for multiple choice [−1.51, 3.28] (p = .46). These no-block models do not decompose block-adjusted total performance or establish a difference between components. At post-test, 46% reached the multiple-choice maximum.

Council-scored relational-quality shares on written responses rose from 8% to 49% with tutor access and from 8% to 27% in holdout (Figure 8). The adjusted depth coefficients are positive (Table 8); under the post hoc Holm correction both depth endpoints fall at p = .058. Online Appendix Section D reports provider-panel sensitivities and Section H the multiplicity calculation.

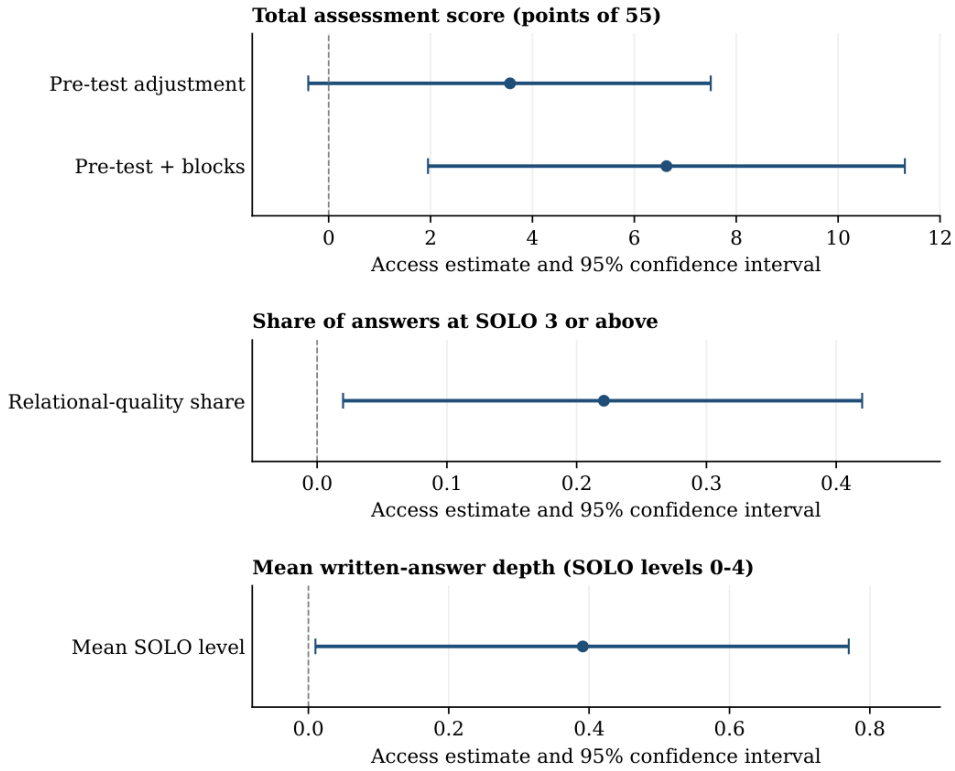


Figure 7: Access estimates and 95% intervals on their original outcome scales. Total-score estimates vary with adjustment; the written-depth estimates have post hoc Holm-adjusted p-values of .058 and .058.

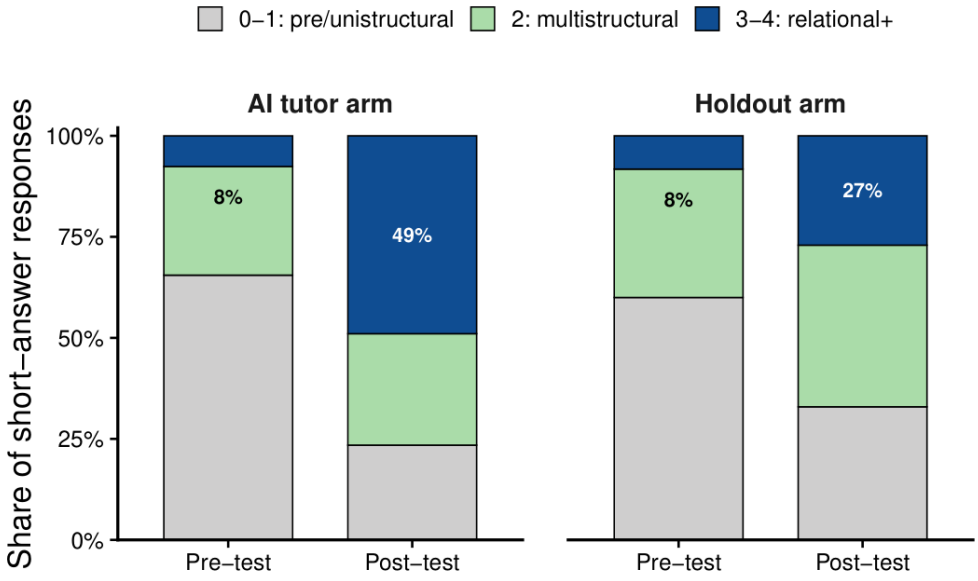


Figure 8: Council-scored written-answer SOLO distributions by arm and administration. The share reaching relational quality or above increased from 8% to 49% in the tutor arm and from 8% to 27% in holdout.

Turning to modality, the directly randomized Weeks 8–9 contrast, the prioritized H2 estimate, was −.079 points out of 11 (95% CI [−1.103, +.944], p = .875; 29 students). The five-week platform-assignment

panel gave −.006 (95% CI [−.616, .604], p = .985), and the attended-weeks and all-week-calendar panels were likewise near zero (Online Appendix Table S1; Figure 9).

No panel detects a voice advantage. The prioritized Weeks 8–9 estimate is imprecise: its interval admits differences of about one point of 11 in either direction. The weekly SOLO contrast also showed no detected difference (p = .99).

Figure 9 displays the four modality contrasts.

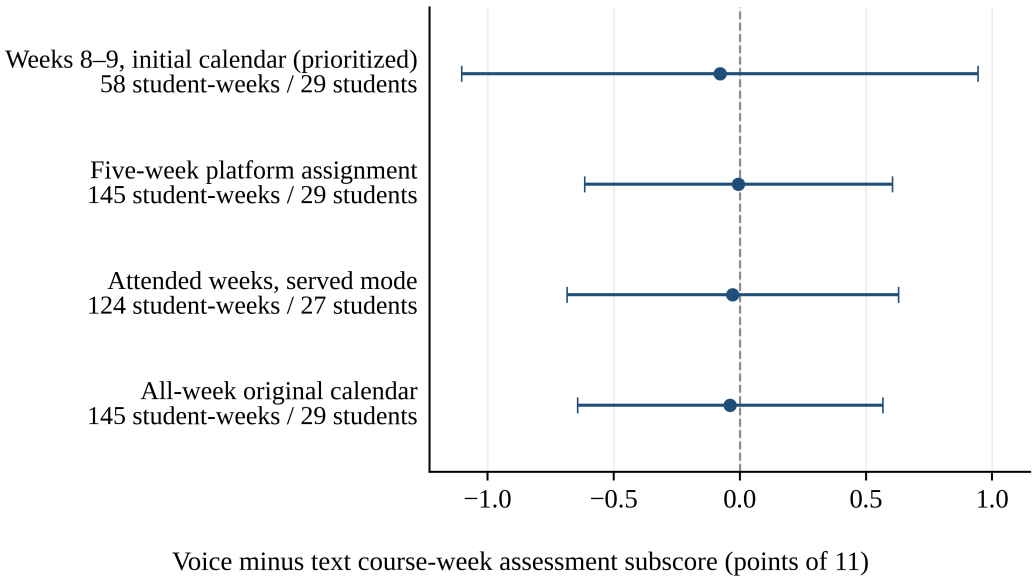


Figure 9: Voice-minus-text course-week assessment contrasts on the 0–11 scale. Points and bars show estimates and 95% intervals. Weeks 8–9 preserve the initial randomized calendar comparison and are the prioritized estimate; later platform assignment depends on prior completion.

In sum, access estimates are positive and specification-sensitive; H2 is not supported; and the exploratory H3 interaction between access and item cognitive demand is not significant.

4.6. Exploratory associations

Within the tutor arm, each additional completed week was associated with +2.00 points of gain (95% CI [.37, 3.6], raw p = .018). This H4 coefficient describes self-selected completion rather than a causal dose response.

The within-student association between turn density and course-week subscore was −.20 points per additional turn per minute (p = .63). Across students, voice–text differences in density and subscore correlated at r = −.16 (p = .41). More turns did not predict higher weekly scores, within or across students: dialogue volume is not a proxy for learning in this setting.

4.7. Convergent evidence from the instructor's midterm and final

Both course-total models use all 86 students (Table 9). The midterm access contrast was +1.32 percentage points (HC1 95% CI [−1.91, 4.55], p = .418); the final was +2.57 [.12, 5.01] (p = .040). Final HC3 inference gives p = .068. Among the 46 retained study-assessment students, the final estimate was +2.53 [−1.27, 6.34] (HC1 p = .18). The full-cohort final broadens outcome coverage on a distinct instrument.

Table 9: Instructor-course midterm and final access contrasts

Instrument and sample	Estimate	95% interval	Raw p
Midterm /100, all assigned (n = 86)	+1.32	[-1.91, +4.55]	.42
Final /100, all assigned (n = 86)	+2.57	[+.12, +5.01]	.040
Final /100, retained sample (n = 46)	+2.53	[-1.27, +6.34]	.18

Note. Exploratory ANCOVAs adjust for workbook study pre-test and ability blocks, with HC1 errors. Both instruments are normalized to 0–100; estimates are percentage points. Full-cohort and retained final samples overlap.

The course-final result is directionally consistent with the positive study-assessment access estimates and provides convergent evidence on a distinct instrument. The weekly-check contrast was +.91 (p = .45). Item-level analyses yielded a final access-by-higher-order interaction of +8.91 points of item credit (95% CI [.97, 16.86], raw p = .028, secondary-family q = .084), compared with a midterm interaction of -3.89 (95% CI [-8.65, .86], p = .11). The final interaction rests on two higher-order items and remains exploratory (Online Appendix Table S2).

4.8. Observed delivery expenditure

Field-period provider billing totaled $562.59: $413.46 for voice and $149.13 for text, a total-bill ratio of 2.77. Recorded billing volumes were 1,880 connected voice minutes and 9,038 text messages, averaging $.220 per minute and $.0165 per billed message. Billed text messages use a different unit from the 1,705 student text messages in the dialogue corpus.

Table 10 reports billing units and totals. Across both modes, expenditure was $10.82 per offered tutor seat, $14.81 per active user, and $3.91 per completed student-week, using 52 seats, 38 active users, and 144 completions.

Table 10: Observed delivery expenditure

Cost measure	Voice	Text
Billing unit	Connected minute	Message
Recorded volume	1,880 minutes	9,038 messages
Average billed cost per unit	$0.220	$0.0165
Field-period expenditure	$413.46	$149.13
Weekly expenditure, rounded dollars	64 / 94 / 91 / 103 / 62	28 / 28 / 37 / 25 / 31

Note. Total expenditure was $562.59, including $390.88 in recorded LLM charges (69.5%). Expenditure per offered seat, active user, and completed student-week was $10.82, $14.81, and $3.91 across both modes.

These amounts exclude development, integration, instructor labor, and institutional overhead. The 2.77 total-bill ratio compares observed bills at different exposures.

5. Discussion

5.1. Summary of findings against the competing accounts

We organize discussion around three main findings. First, offering the structured AI tutor was associated with better performance on the study post-test, greater written-answer depth, and a higher instructor's final across the full cohort, with no detected midterm difference. Gain per completed week also increased. H1, H3 and H4 are supported. Second, voice changed the form of student practice: sessions had roughly twice the turn density, more than twice the accumulated questions, and longer messages, with similar engaged time. Third, we detected no voice advantage on weekly assessments in the directly randomized contrast or the sensitivity panels, and more voice turns relative to text did not predict higher scores within or across students. H2 was not supported.

Read against Table 1, the modality findings are consistent with text facilitation, possibly reinforced by the shared visual channel. Social agency predicts that a speaking tutor raises engagement and, through it, learning; the participation half appeared, while the assessment advantage was not detected. The production effect predicts a memory advantage for content students voiced; none was detected, although the weekly subscores are a coarse test of it. The effort to produce text predicts longer formulation pauses and less interaction frequency compared to voice interaction yet does not create a voice learning advantage. The pattern resembles that found by Litman et al. (2006) and D'Mello et al. (2011) with constrained dialogue systems, now in a generative AI tutor used repeatedly over a course.

5.2. Interpreting the access effect

The treated condition evaluates tutor availability, structured activity, feedback, and incentives alongside existing course resources. Additional practice is a plausible explanation for the positive estimates. The written-depth findings are consistent with explanation and justification. The instructor’s final and midterm supply distinct course instruments. Control treatments that actively promote practice and give feedback independent of AI would test which configuration elements contribute most to learning outcomes.

5.3. Interpreting the dialogue and assessment pattern

Recorded participation differed in rhythm and extent between voice and text. Greater turn density, more questions, and longer messages show increased participation. Whether that participation involved constructive reasoning depends on the content of the contributions and the feedback that followed; voice transcripts also contain fillers, restarts, and spoken elaboration that inflate word counts without adding content. ICAP research directs attention to explanations, connections, and revisions as evidence of constructive activity that might differ between voice and text (Chi et al., 1994; Chi & Wylie, 2014).

Three features of the design plausibly explain the dialogue–assessment pattern where we observed no voice advantage. First, typing encourages ideation before submission, while speech supports immediate response production (D'Mello et al., 2011): the median typed reply followed a 26-second pre-keystroke pause, whereas playback-adjusted spoken replies followed the tutor's turn within a few seconds. If composing an answer supports learning, text provides a formulation opportunity whose contribution may differ from spontaneous speech. Second, both modes supplied readable explanations through the transcript and blackboard, so visual support relevant to multimedia modality effects (Ginns, 2005; Mayer et al., 2001) was held constant rather than varied. Third, the language model received transcripts rather

than audio and could not respond directly to intonation or affect. These are plausible readings of the design; the process–score associations do not identify which mechanism applies.

5.4. Preference, completion, and sustained voice use

Student preferences moved toward voice through Week 11, then declined. Final-week completion was 9/12 voice opportunities and 17/18 text opportunities under completion-dependent assignment. Quiet-space requirements, survey nonresponse, or novelty could explain this divergence; the data do not distinguish them. Preference and sustained use warrant separate measurement.

6. Implications for research and practice

For instructors and instructional designers, the practical lesson is to evaluate the structured practice offered by a tutor separately from its interaction channel. A tutor that requires an attempt before explaining, asks one question at a time, and ends with a transfer task can be configured in either mode. This study detected no assessment advantage from adding voice. Designers can specify attempts, justifications, graduated hints, and application checks, inspect dialogue against those requirements, and align assessments with the reasoning the tutor asks for. A workflow comparison and a fidelity audit would establish which elements carry the access effect.

For institutions deciding whether to pay for voice, the relevant comparison combines assessed performance, participation, accessibility, and cost. At this deployment's prices and exposures, voice billing was 2.77 times text billing, with roughly double the dialogue density, similar engaged time, and no detected assessment advantage. Institutions can offer voice as an option where it serves accessibility or engagement needs rather than assume a test-score return from making it the default. Those potential benefits require direct evaluation; the final-week completion pattern also suggests monitoring sustained use rather than inferring it from early preference.

For researchers and developers, the shared transcript and blackboard are features to vary next: a voice-only arm without visual support would test whether the shared channel attenuates modality differences. Native speech-to-speech models that process audio directly would test whether responding to intonation and affect changes the comparison. Evaluation should pair such contrasts with human-coded explanation quality, continued use without incentives, and billing aligned with exposure.

7. Limitations and Future Research

First, selection and estimation: 46 of 86 students entered the study-assessment analysis, with higher baseline performance among retained students in both arms. IPW requires correctly modeled observation probabilities and conditional exchangeability; reported HC1 errors hold estimated weights fixed. Lee point bounds require monotonic retention and are neither full-cohort nor block-ANCOVA bounds. Approximate imputation uses different baseline scoring, block conventions, and donors from the primary analysis; both full-cohort total-score intervals cross zero. Sparse blocks and covariance choice affect inference.

Second, measurement: short-answer scores rest on automated grading, and tutor feedback on response structure could favor the format those graders reward. Provider agreement does not detect a shared grading bias. Multiple-choice performance is constrained by a post-test ceiling, three-item course-week subscores provide limited coverage, and variant difficulty and balance were not independently verified.

The process classifier has not been validated against human coding, reply intervals include different activities in each mode, and tutor playback is estimated.

Third, the modality comparison: the initial sequence was randomized, but assignment depended on completion from Week 10, and alternation, review, and preview permit carryover across topics. The results therefore describe weekly differences within an alternating regime, not sustained voice-only or text-only tutoring, and the prioritized Weeks 8–9 estimate is imprecise. The shorter-window priority reflects the present analysis; its pre-specification is not documented.

Fourth, scope: the study concerns one voluntary, incentivized graduate-course deployment using a transcript-based speech pipeline. Demographic responses were unavailable, preventing composition, coverage, and arm-balance assessment. Instructional fidelity, finance-content accuracy, ASR accuracy, and end-to-end latency were not audited. Timing and volume do not measure reflection, cognitive load, social presence, or affect.

The take-home final was unsupervised and showed a ceiling. It does not isolate unaided retention or long-term follow-up. The midterm occurred during intervention but covered course material from the first half not covered by the tutor. Final inference depends on covariance choice. Costs exclude implementation and instructional labor.

Future studies should use active-practice controls, registered assessment rules, direct fidelity measures, and sustained-mode comparisons. Order-sensitive analyses can separate sequence and carryover; broader institutional coverage and attrition-aware power calculations can establish when voice justifies its cost.

8. Conclusion

Offering the structured AI tutor was associated with higher post-test scores and greater council-scored written-answer depth, with a positive access contrast on the instructor's final across all 86 assigned students. The retained study-assessment estimates vary with specification; the full-cohort final provides convergent evidence from a distinct instrument. Voice sessions had roughly twice the dialogue density of text sessions and more question-asking, while no weekly assessment advantage was detected and dialogue volume did not predict performance. Pooled expenditure was $3.91 per completed student-week, and observed voice billing was 2.77 times text billing.

The findings are consistent with benefits from offering the structured practice bundle, while the modality pattern extends earlier dialogue-tutoring evidence to a generative AI tutor. Institutions can offer voice where it serves an evaluated access or engagement need, without assuming additional learning gains. An incentivized active practice control would test the contribution of instructional configuration, and a speech-native tutor that processes audio would test whether the modality pattern changes.

Declarations

Ethics approval and informed consent: Institutional review board approval was granted on June 19, 2026 (institution and protocol number withheld for anonymous review). Participants gave informed consent before randomization; consent covered stored transcripts, voice-session audio, blackboard content, interaction telemetry, assessment responses, retention, access controls, and third-party processing.

Funding and competing interests: Study support is disclosed on the separate title page, with identifying details withheld for anonymous review. Apart from that support, no additional funding or support was

received. The authors declare no other competing interests; the infrastructure provider had no role in study design, analysis, or reporting.

Author contributions: The named author contribution statement is provided on the separate title page.

Data availability: Student-level data are not publicly available because they contain protected student information that are subject to student privacy and FERPA restrictions.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the authors used GPT-6.1-sol (OpenAI) to assist with restructuring the manuscript and Codex (OpenAI) to assist with the code that produced the plots and diagrams in Figures 1, 3, 7, and 9 and Online Appendix Figure S1. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Agrawal, K., Athey, S., Kanodia, A., Nath, S., & Palikot, E. (2026). *The economics of algorithmic personalization: Evidence from an educational technology platform*. NBER Working Paper No. 34950. https://doi.org/10.3386/w34950

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman. https://www.loc.gov/item/00063423/

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology, 30* (1), 117–139. https://doi.org/10.1016/j.cedpsych.2004.07.001

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40* (4), 199–209. https://doi.org/10.1207/s15326985ep4004_2

Azevedo, R., & Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition—Implications for the design of computer-based scaffolds. *Instructional Science, 33* (5–6), 367–379. https://doi.org/10.1007/s11251-005-1272-9

Bapna, R., Bapna, S., Kumar, P., Pamuru, V., Umyarov, A., & Yang, M. (2025). *Agentic AI and managers' analytics capabilities: An exploration*. Working paper. Indian School of Business. https://doi.org/10.2139/ssrn.5293722

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

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. Academic Press. https://doi.org/10.1016/C2013-0-10375-3

Bloom, B. S. (Ed.). (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York, NY: David McKay Company.

Chen, Q., Gong, Y. Y., Keil, M., Liu, S., & Lu, Y. (2024). Conceptualization and measurement of voice-interaction usability: The development of cooperative principle theory for smart product use. *MIS Quarterly, 48* (3), 1009–1046. https://doi.org/10.25300/MISQ/2023/17525

Chi, M. T. H. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10* (7), S33– S49. https://doi.org/10.1002/(SICI)1099-0720(199611)10:7<33::AID-ACP436>3.0.CO;2-E

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1* (1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18* (3), 439–477. https://doi.org/10.1207/s15516709cog1803_3

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49* (4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education, 210*, Article 104967. https://doi.org/10.1016/j.compedu.2023.104967

D'Mello, S. K., Dowell, N., & Graesser, A. (2011). Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied, 17* (1), 1–17. https://doi.org/10.1037/a0022674

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114* (4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Fernandes, T., & Oliveira, E. (2021). Understanding consumers' acceptance of automated technologies in service encounters: Drivers of digital voice assistants adoption. *Journal of Business Research, 122*, 180– 191. https://doi.org/10.1016/j.jbusres.2020.08.058

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General, 144* (3), 674–687. https://doi.org/10.1037/xge0000070

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15* (1), Article 6. https://doi.org/10.3390/soc15010006

Ginns, P. (2005). Meta-analysis of the modality effect. *Learning and Instruction, 15* (4), 313–331. https://doi.org/10.1016/j.learninstruc.2005.07.001

Hoogerheide, V., Deijkers, L., Loyens, S. M. M., Heijltjes, A., & van Gog, T. (2016). Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them. *Contemporary Educational Psychology, 44–45*, 95–106. https://doi.org/10.1016/j.cedpsych.2016.02.005

Jacobs, S., & Kiesler, N. (2025). GenAI voice mode in programming education. In *Proceedings of the 25th Koli Calling International Conference on Computing Education Research* (pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3769994.3770001

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13* (4), 351–371. https://doi.org/10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331* (6018), 772–775. https://doi.org/10.1126/science.1199327

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

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19* (3), 239–264. https://doi.org/10.1007/s10648-007-9049-0

Koedinger, K. R., Kim, J., Jia, J. Z., McLaughlin, E. A., & Bier, N. L. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. In *Proceedings of the Second ACM Conference on Learning @ Scale* (pp. 111–120). https://doi.org/10.1145/2724660.2724681

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*. arXiv preprint. https://doi.org/10.48550/arXiv.2506.08872

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86* (1), 42–78. https://doi.org/10.3102/0034654315581420

Lachner, A., Ly, K.-T., & Nückles, M. (2018). Providing written or oral explanations? Differential effects of the modality of explaining on students' conceptual learning and transfer. *The Journal of Experimental Education, 86* (3), 344–361. https://doi.org/10.1080/00220973.2017.1363691

LearnLM Team. (2024). *LearnLM: Improving Gemini for learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2412.16429 Version cited: https://arxiv.org/abs/2412.16429v1

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies, 76*(3), 1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence, 6*, Article 100213. https://doi.org/10.1016/j.caeai.2024.100213

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Article 1121, 1–22. https://doi.org/10.1145/3706598.3713778

Litman, D. J., Rosé, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education, 16* (2), 145–170. https://doi.org/10.3233/IRG-2006-16(2)04

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106* (4), 901–918. https://doi.org/10.1037/a00371

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26* (4), 390–395. https://doi.org/10.1177/0963721417691356

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36* (3), 671–685. https://doi.org/10.1037/a0018785

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93* (1), 187–198. https://doi.org/10.1037/0022-0663.93.1.187

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95* (2), 419–425. https://doi.org/10.1037/0022-0663.95.2.419

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology, 72*, 609–633. https://doi.org/10.1146/annurev-psych-010419-051019

McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa… Examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior, 99*, 28–37. https://doi.org/10.1016/j.chb.2019.05.009

Mollick, E. R., Mollick, L., Bach, N., Ciccarelli, L. J., Przystanski, B., & Ravipinto, D. (2024). *AI agents and education: Simulated practice at scale*. arXiv preprint. https://doi.org/10.48550/arXiv.2407.12796

Moreno, R., & Mayer, R. E. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology, 92* (4), 724–733. https://doi.org/10.1037/0022-0663.92.4.724

Moriuchi, E. (2019). Okay, Google!: An empirical study on voice assistants on consumer engagement and loyalty. *Psychology & Marketing, 36* (5), 489–501. https://doi.org/10.1002/mar.21192

Moriuchi, E. (2021). An empirical study on anthropomorphism and engagement with disembodied AIs and consumers' re-use behavior. *Psychology & Marketing, 38* (1), 21–42. https://doi.org/10.1002/mar.21407

Nass, C., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56* (1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). https://doi.org/10.1145/191666.191703

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition, 40* (3), 326–338. https://doi.org/10.3758/s13421-011-0165-1

Palmer, K. (2026, October 7). A university's research on its own AI tool raised more questions than answers. *Inside Higher Ed*. https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2026/10/07/maryland-ai-research-raised-more-questions

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. *PLOS ONE, 19* (5), Article e0304013. https://doi.org/10.1371/journal.pone.0304013

Qin, F., Hao, Z., Yu, J., Liu, Z., & Zhang, Y. (2025). *AI instructional agent improves student's perceived learner control and learning outcome: Empirical evidence from a randomized controlled trial*. arXiv preprint. https://doi.org/10.48550/arXiv.2505.22526

Qiu, L., & Benbasat, I. (2009). Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of Management Information Systems, 25* (4), 145–182. https://doi.org/10.2753/MIS0742-1222250405

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Reich, J., & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science, 363* (6423), 130–131. https://doi.org/10.1126/science.aav7958

Riegel, K., Hua, Y. C., Denny, P., Pădurean, V.-A., Leinonen, J., Prather, J., & Singla, A. (2026). Say what? Examining text and voice input modalities for prompt-based programming in computing education. In *Proceedings of the 31st ACM Conference on Innovation and Technology in Computer Science Education V. 1 (ITiCSE 2026)* (pp. 252–258). Association for Computing Machinery. https://doi.org/10.1145/3803400.3809397

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20* (9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

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

Rzepka, C., Berger, B., & Hess, T. (2022). Voice assistant vs. chatbot—Examining the fit between conversational agents' interaction modalities and information search tasks. *Information Systems Frontiers, 24* (3), 839–856. https://doi.org/10.1007/s10796-021-10226-5

Schroeder, J., & Epley, N. (2015). The sound of intellect: Speech reveals a thoughtful mind, increasing a job candidate's appeal. *Psychological Science, 26* (6), 877–891. https://doi.org/10.1177/0956797615572906

Schroeder, J., & Epley, N. (2016). Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General, 145* (11), 1427–1437. https://doi.org/10.1037/xge0000214

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2021). Voice in human-agent interaction: A survey. *ACM Computing Surveys, 54* (4), Article 81, 1–43. https://doi.org/10.1145/3386867

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2021). Texting with humanlike conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems, 22* (4), 931–967. https://doi.org/10.17705/1jais.00685

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333* (6043), 776–778. https://doi.org/10.1126/science.1207745

Stromberg, D., Lei, V., & Wu, Y. (2026). *The generative AI learning penalty: Evidence from Chinese secondary education*. Working paper. SSRN. https://doi.org/10.2139/ssrn.6868618

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46* (4), 197–221. https://doi.org/10.1080/00461520.2011.611369

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21* (3), 209–249. https://doi.org/10.1207/S1532690XCI2103_01

Wagner, K., Nimmermann, F., & Schramm-Klein, H. (2019). Is it human? The role of anthropomorphism as a driver for the successful acceptance of digital voice assistants. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2019.169

Wiles, E., Krayer, L., Abbadi, M., Awasthi, U., Kennedy, R., Mishkin, P., Sack, D., & Candelon, F. (2026). Generative AI and the temporary upskilling of knowledge workers. *Nature Human Behaviour*. In press. https://emmawiles.github.io/storage/reskill.pdf

Xu, Y., Aubele, J., Vigil, V., Bustamante, A. S., Kim, Y.-S., & Warschauer, M. (2022b). Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Development, 93* (2), e149–e167. https://doi.org/10.1111/cdev.13708

Xu, Y., Vigil, V., Bustamante, A. S., & Warschauer, M. (2022a). Contingent interaction with a television character promotes children's science learning and engagement. *Journal of Applied Developmental Psychology, 81*, Article 101439. https://doi.org/10.1016/j.appdev.2022.101439

Zhai, X., Haudek, K. C., Stuhlsatz, M. A. M., & Wilson, C. (2020). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation, 67*, Article 100916. https://doi.org/10.1016/j.stueduc.2020.100916