ACADEMIA International Journal for Social Sciences
Volume 4, Issue 3, 2025 ISSN-L (Online): 3006-6638

Enhancing Learning Outcomes through AI-Based Tutoring Systems: A Study on Student Motivation and Academic Achievement

Zeeshan Pervaiz
dubai.intpsc@gmail.com
PhD Education. Hazara University Mansehra, Pakistan

Abu Bakar
Abubakarrizvi786@gmail.com
Muslim Youth University Islamabad, Pakistan

Sohail Saddique
sohailsaddique618@gmail.com
Muslim Youth University Islamabad, Pakistan
Corresponding Author: * Zeeshan Pervaiz dubai.intpsc@gmail.com

Received: 20-07-2025 **Revised:** 24-08-2025 **Accepted:** 06-09-2025 **Published:** 18-09-2025

ABSTRACT

Purpose: To determine whether an artificial intelligence (AI)-based tutoring system (AITS) is more effective in terms of academic success and motivation, as well as to investigate causative influences of motivation. Techniques: It was a pre-registered randomised trial in 24 classes (N=602; Grade 7-10), with assignment to AITS or business-as-usual either at the student or class level. The intervention provided adaptive sequencing, stepwise feedback, mastery thresholds, and spaced review in 8-12 weeks. The outcome measures included Post-test achievement that was curriculum-based; Intrinsic Motivation Inventory and MSLQ subscales were the secondary outcome measures.
The ANCOVA and multiple imputation linear mixed models were analysed and then multilevel mediation and moderation followed. Findings: AITS brought about a 5.1-point (d[?]0.40; p<.001) posttest-controlling effect. Interest/enjoyment and perceived competence went up (d=.20-.45). The achievement effect was mediated by interest ≈24%. The effects were greater in students with lower baseline scores and rose with usage (to approximately 12 hours) after which it was diminishing in returns. Results were strong to clustering correction, different scaling and sensitivity tests. Conclusions: Under normal classroom time, AITS has the potential to improve performance through the improvement of motivational states and effective engagement, especially with occurrence in lower-baselin learners. It should be implemented to focus on minimum viable dosage, step-level feedback of high quality, and analytics-informed coaching. The future work must also be cross-subject and cross-term, include retention results, and have equity and cost-effectiveness audits. All materials will be shared.
Keywords: *AI tutoring; adaptive learning; mastery learning; student motivation; self-determination theory; learning analytics; randomized controlled trial.*

INTRODUCTION

Individual tutoring systems based on artificial intelligence (AI) include intelligent tutoring systems (ITS) and more modern K-12 and higher education tutors with generative-AI-enhanced tutoring; they are rapidly becoming embedded in systems. Several decades of research in learner modeling, adaptive feedback, and progressive scaffolding of problems have evolved into scalable platforms that can offer personalized practice and real-time hints at only a fraction of the cost of human tutoring (Wang et al., 2023; Son, 2024). The recent massive consumption statistics indicate that properly designed ITS can generate quantifiable increase in accomplishment, even amidst unfavourable academic circumstances. To illustrate, when millions of mathematics exercises are analysed, positive growth trajectories are found in cases of long-term ITS activities, and their patterns depend on the initial level of proficiency (Spitzer and Moeller, 2023; Spitzer and Moeller, 2024). Simultaneously, the emergence of conversational generative systems, which are applied as on-command explainer, Socratic guide, or feedback engine, have led to another round of tests and syntheses in which better results in performance are recorded, along with affective and engagement outcomes (Deng et al., 2024; Heung and Chiu, 2025; Lo et al., 2024).

Irrespective of this development, the evidence on the use of AI tutoring to enhance learning is still disproportional. In turn, previous syntheses mention that most of the results are concerned with achievement, whereas non-cognitive results (e.g., intrinsic motivation, engagement) are under-investigated or quantified heterogeneously (Wang et al., 2023; Son, 2024). Simultaneously, the generative-AI canons are growing at a phenomenally fast rate yet differ in the way it is designed, in which its comparison is conducted, and reporting criteria are applied, which makes it challenging to attribute causality to the AI or the pedagogy the AI produces (Deng et al., 2024). This has led to the fact that the academic community is yet to rigorously and comprehensively test whether academic performance and student motivation both improve with AI tutoring, and whether motivational improvement contributes to performance improvement.

Traditional forms of tutoring, including peer tutoring, office hours, small-group large-scale tutoring, are resource-intensive, hard to differentiate, and lacking in quality. Schools that face huge populations of students with varying readiness levels, frequently are unable to provide enough responsive feedback or practise opportunity when and where it is required. Scalable, always-available supports offered by AI-based tutors are claimed to be able to scale, hints and pacing to individuals, though, concrete evidence that captures the joint effects of achievement and motivation as well as the mechanisms and moderators of these effects are scarce. This research fills that gap by testing the hypothesis of more post-test performance and student motivation with an AI tutor learning compared to business-as-usual learning, as well as testing the motivational mediation hypothesis and usage/moderation hypothesis.

Our research is based on Self-Determination Theory (SDT) and Expectancy-Value theory (EVT). According to SDT, high-quality motivation and persistence among learners are enhanced through an instruction that encourages autonomy (meaningful choice and rationale), competence (challenging tasks preferably with informative feedback), and relatedness (warmth, care, and belonging). The meta-analytic findings indicate that the need-supportive teaching is closely linked to satisfying the basic needs and, consequently, the well-being and performance (Slemp et al., 2024). In addition, interventions based on SDT in education have proven to have stable motivational effects and in other studies even downstream performance benefits - in line with the concept that the instruction design that influences need support can modulate engagement and learning (Wang et al., 2024). Adaptive feedback, instant knowledge of results, and step-by-step scaffold designs can be designed to provide these need supports at scale, only with AI tutors (Wang et al., 2023; Son, 2024).

EVT (since frequently interpreted as Situated Expectancy-Value Theory, SEVT) argues that achievement behaviour is shaped by expectancies of success and value of the tasks (interest, utility, attainment) to them in situational context (Eccles and Wigfield, 2024). Classroom-based reviews report the effects of feature of everyday instruction (e.g. relevance cues, success-structure, cognitive load) to these perceptions, whose downstream effects are engagement and persistence (Tang et al., 2022). Combining SDT and EVT, one can posit that an autonomy/competence/relatedness-based design aspect will increase intrinsic motivation and engagement; with gaining expectancies and values, students are expected to invest an effective effort, resulting in greater achievement. This motivational pathway is explicitly tested by our study and moderate factors (e.g., prior achievement, intensity of treatment use) proposed by previous ITS studies are also considered (Spitzer and Moeller, 2023).

Empirical. Our pre-registered, curriculum-congruent experiment is a comparison of AI tutoring to business-as-usual with cognitive (achievement) and non-cognitive (motivation/engagement) endpoints, which bridges a gap in the literature (Wang et al., 2023). Theoretical. We operationalize SDT and EVT in an AI-tutoring scenario jointly and find out motivational mediation and usage/prior-achievement moderation. Practical. Design (e.g., autonomy-supportive prompts, competence-calibrated scaffolds, socially-attuned feedback) and implementation levers that practitioners can utilize to enable maximum learning and motivation are identified by us, which are consistent with the emerging evidence on ITS and generative-AI deployments (Deng et al., 2024; Heung and Chiu, 2025; Lo et al., 2024).

Overall, conventional tutoring is impossible to scale to suit various demands, whereas AI tutoring is promising, although little evidence demonstrating that it can improve both academic performance and student motivation simultaneously and the mechanisms connecting them is available. The research question of this paper is to determine, therefore, whether an AI tutor results in better achievement and motivation relative to business-as-usual, whether motivational change mediates the achievement effects, and whether achievement effects are moderated by previous achievement, or the intensity of usage. We plan to accomplish (a) estimating the causal impact of AI tutoring on post-test performance with baseline correction; (b) testing differences in intrinsic motivation/engagement; (c) testing a theory-based mediation pathway based on SDT/EVT; (d) exploring sub-group differences based on their prior achievement and usage to inform targeting and implementation.

LITERATURE REVIEW

Definitions and core capabilities. Intelligent tutoring systems (also known as AI-based tutoring systems, AITS) imitate certain aspects of one-to-one human tutoring, keeping a model of a learner, customising tasks, and giving them personalised advice in real time. The methods that are combined in the contemporary AITS include knowledge tracing, mastery learning pathways, and natural-language dialogue to provide adaptive sequences of practise, immediate and detailed feedback, demand-based hints, and automated worked examples. They are now incorporating generative AI to aid in rich explanations and open-ended problem solving and capture detailed interaction data to be used in analytics (Lin et al., 2023; Son et al., 2024).

Progression in the mastery and adaptability. Progression Mastery Mastery based progression maps performance of students to fine grained skill models, when there is weak mastery evidence, remedial content is selected by the system in an adjustive fashion, and it is spaced over time to encourage enduring retention. The adaptive task choice is based on the idea of balancing between difficulty and success through learner modeling (e.g., Bayesian or deep models) to optimize time-on-task and reduce frustration (Conati et al., 2021; Lin et al., 2023). Feedback and hints. AITS provide task-contingent feedback at rapidity in nature that is as fine as a signal of correctness, or as a hint that is given step by step. Explanations can also be personalised (why a hint was provided, what misunderstanding is probable) to boost trust, interest, and learning, particularly with learners who do not use feedback effectively in general (Conati et al., 2021).

R space on review and retrieval. A lot of AITS review with spacing algorithms, interweaving acquired skills previously learned, to overcome forgetting. Spaced retrieval and cumulative review are administered in the form of periodic mixed-skill practice sets or resurgence knowledge checks, which is frequently elicited by estimates of the forgetting curve (Kim and Webb, 2022; Bego et al., 2024). What are the recent evaluations? AITS usually exhibit small-to-moderate positive effect on achievement as compared to business-as-usual instruction with the heterogeneity being attributed to subject, implementation fidelity and outcome type. A massive systematic review (2011-2022) of studies (field RCTs/ quasi-experiments) of the type of systematic review included only studies employing ITS in authentic scenarios, which showed that ITS produces benefits, albeit with a disproportionate level of methodological rigor and reporting (Wang et al., 2023).

Effect sizes and exemplars. Reviews in mathematics have recently summarised common effects in the g ≈ 20-40 range of achievement tests when AITS are in curricula (Son et al., 2024; Letourneau et al., 2025). New randomized trials involving LLM-enhanced tutors demonstrate greater short-term benefits in certain college settings: a 2025 randomized trial found that students using an AI tutor performed better on standardized tests than their counterparts in areas with active learning (Kestin et al., 2025). Such findings are promising but they need to be replicated at levels, subjects and longer follow-ups.

Methodological gaps. In the literature, various weaknesses are repeated: (a) the lack of pre-registration and attrition reporting; (b) the clustering effect is not noted, where analysis is done (students in a class/school), which inflates the precision; (c) there is a problem with the alignment of outcomes (locally developed tests against standardized tests); and (d) there are weak checkpoints on implementation faithfulness and instructor effects (Wang et al., 2023; Letourneau et al., 2025). Mediation of effects through motivation/ engagement has rarely been studied through rigorous causal models, and hence the how of AITS efficacy has been under-specified.

AI-based tutoring systems (AITS) may have a plausible effect on achievement via motivation. By influencing what the learners focus on, the duration of focus, and the perceptions of achievement or failure, AITS can change the motivational states which, in its turn, is reflected in the enhanced performance. It has three design levers, specifically feedback timing and quality, goal setting, and autonomy support, which are particularly relevant in digital environments, where the interactions are frequent and data intensive.

First, it is important in terms of feedback timing and quality. The meta-analytic evidence demonstrates that feedback is a reliable way to enhance the learning process, and the impact of these messages is moderated by the specificity, task-focus, and actionability of messages; in the digital environment, properly developed feedback interventions may convincingly produce effects of approximately $d \approx 40$ (Wisniewski et al., 2020; Brummer et al., 2024). Step-level explanations will be able to maintain the interest with the immediate explanations that will help to bridge the gaps in knowledge when they appear and provide the next step. On the other hand, delayed feedback may lead to reflection and diagnosis of errors. Recent studies indicate that the optimal time can be determined based on the demands of the task and the features of the learner-e.g., novices can receive more extensive benefits with the help of quick scaffolds, and advanced learners can receive benefits with delays that promote self-explanation (Ryan et al., 2024).

Second, in AITS goal setting acts as a self-regulatory scaffold. Cues to make specific, difficult and attainable goals it is combined with observable indicators of progress can boost planned studying time, as well as assist students to regulate effort. Digital course syntheses show a positive impact on performance and psychological outcomes, but the size of this effect depends on the nature of the goal (process or outcome), the fineness of progress feedback, and the context of learning (Williamson et al., 2024). In reality, weekly SMART targets, micro-goals based on the mastery levels, and nudges to show the discrepancies in progress are practical applications.

Third, the support of autonomy is a key to maintaining persistence. Interfaces that bring about significant options (e.g., selecting the next sub-goal, choosing the hint depth, or postponing a review), give justifiations to tasks, and recognise difficulty are consistent with the self-determination theory and are likely to enhance perceived autonomy and competence. These can turn students who are compliant to those who are endorsing so that practise becomes self-initiated but not forced. Despite the fact that a large number of formative measurement tools have been developed before 2020, their validation in the context of online and professional learning remains valid to date, which allows assessing the changes in autonomy, competence, and interest during AITS consumption credibly.

These levers contribute to a logical playbook: provide concrete, task-oriented feedback within the appropriate timeframe to the learner; incorporate formal, phase-based goal setting; and develop autonomy-supportive options that have clear rationale. They have been shown to establish motivational conditions that increase the likelihood of AITS to transform time-on-task into lasting learning benefits in situations where they are implemented together. Measures used. IMI (Intrinsic Motivation Inventory) and MSLQ (Motivated Strategies for Learning Questionnaire) are the most frequently used measures of motivation and self-regulated learning. They are structurally valid and reliable in a digital and health-profession setting as recently validated by researchers (Cook and Skrupky, 2025).

Behavioural involvement as an intermediate product. Leading indicators of eventual performance in AITS are clickstream indicators: time-on-task, regularity of the session, hint requests, and persistence during an on-task. Indicatively, research has identified a positive correlation between stable platform time and regularity of study rhythms and high course grades, despite the control of what they initially believed (Roh et al., 2021). With the maturity of learning analytics, process based experiments demonstrate that more informative, usage-adaptive feedback can lead to improved perceived helpfulness and self-regulation, which depend on feedback literacy (Learning Environments Research meta-findings; Educational Technology & Society/ET&S and ET&S derivatives).

From data to intervention. Just-in-time supports, which could be a short video, worked example, or a message coach, can be activated by predictive analytics to identify disengagement (e.g., long idle spans, hint was reused multiple times). A higher educational study on learning-analytics demonstrated that 2024 time-management trends based on trace data do forecast performance and are subject to nudging (van Sluijs et al., 2024). Meta-analytic data regarding learning analytics intervention indicate a small-to-moderate improvement but indicate the variability and that they should be meticulously designed (Zheng et al., 2025).

In the recent reviews and tests, the same image appears: AITS positively affect performance on average, and they are likely to effect consumer behaviour through the creation of motivational states (interest, perceived competence, goal commitment) and behaviours (productive time-on-task, help seeking) that are directly related to learning. However, few articles have fitted the causal chain AITS motivation/engagement achievement using contemporary designs.

METHODOLOGY

Design

We will evaluate the causal impact of an AI-based tutoring system (AITS) on student motivation and academic achievement using a **randomized controlled trial (RCT)** as the primary design. Where logistics prevent randomization at the student level, we will employ **cluster randomization** at the classroom level and analyze with multilevel models. If randomization is infeasible at some sites, we will conduct a **pre-registered quasi-experiment** using propensity-score methods and ANCOVA as a complementary design, clearly labeling estimates as non-experimental.

Allocation and Masking.

- *Student-level RCT*: within sections, students are randomly assigned (1:1) to AITS or control using a computer-generated list with blocked randomization on prior achievement (tertiles) and gender to enhance balance.

- *Class-level RCT*: whole classes are randomized (1:1), stratified by teacher and grade. Allocation is concealed until the moment of assignment; outcome assessors are blinded to condition.

Registration & Reporting.

The full protocol (outcomes, models, subgroup plans) will be **pre-registered** on OSF/ClinicalTrials.gov prior to data collection. Reporting will follow **CONSORT-Edu** principles (flow diagrams, balance tables, attrition handling), with a TIDieR-style appendix describing the intervention components and fidelity procedures. We will publish the analysis code and de-identified data dictionary.

Setting & Participants

Context.

The study will take place in public middle and high schools (Grades 7–10) delivering mathematics (primary focus) and science (extension). Schools must have reliable device access (1:1 or cart) and Wi-Fi.

Eligibility.

- *Inclusion (student):* enrolled in participating courses; parental consent and student assent; baseline assessment completed.
- *Exclusion (student):* IEPs requiring alternative assessments that are not compatible with the AITS platform; <50% expected attendance (e.g., long-term leave).
- *Teacher/site:* willingness to implement per protocol; completion of training; agreement to avoid AITS-like tools in control classes.

Sample Size & Power.

Power analyses consider the smallest policy-relevant effect **d = 0.20–0.25** on post-test scores. For **cluster RCTs**, we account for the design effect $\mathbf{DE = 1 + (m - 1) \cdot ICC}$, where m is class size and ICC the intraclass correlation. With baseline covariates (pretest), we apply variance reduction $\mathbf{(1 - R^2\_pre)}$ when estimating required clusters. We plan for ≥80% power at $\alpha = .05$ with two-tailed tests, allowing 15% attrition. Calculations will be verified using Optimal Design/PANGEA and cross-checked by simulation.

Demographics & Baseline Equivalence.

We will collect age, grade, gender, race/ethnicity (as locally categorized), English-learner status, special education status, and free/reduced-price lunch eligibility. Equivalence will be assessed with standardized mean differences ($|SMD| \leq .25$ as acceptable) and covariate-adjusted balance where needed. Any imbalances will be addressed in models (see §3.6).

Intervention (AI tutoring system)

Core Features.

The AITS provides:

1. **Adaptive sequencing** of problems based on a learner model;
2. **Formative feedback** (item- and step-level) with automated explanations and worked examples;
3. **Mastery thresholds** (e.g., ≥0.80 posterior mastery probability or ≥3 consecutive correct at target complexity) to unlock new skills;
4. **Hints on demand** with graduated specificity (general → strategic → bottom-out);
5. **Spaced review** prompts resurfacing previously learned skills.

Implementation & Fidelity.

- *Teacher training:* two 90-minute sessions covering goals, dashboards, troubleshooting, and equity-minded help practices; a quick-reference guide; ongoing office hours.
- *Usage expectations:* 2–3 sessions/week, ~30–40 minutes each, in-class or structured homework; minimum dosage target = 8 hours over 8–12 weeks.
- *Access logistics:* SSO integration; device checks; quiet practice norms; headphones for audio hints.
- *Fidelity monitoring:* (a) system logs (minutes, problems, hint rate, mastery events); (b) monthly teacher check-ins; (c) observation rubric (random 10% of sessions). Thresholds for "adequate fidelity" are pre-specified (e.g., ≥75% of planned minutes, ≥80% of scheduled sessions delivered).

Control Condition.

Business-as-usual (BAU): teacher-selected practice using textbooks/worksheets or a non-adaptive problem set platform without step-level feedback. To minimize **attention/placebo** differences, controls receive comparable teacher-facilitated practice time and conventional feedback routines.

Measures

Academic Achievement (primary).

- *Instrument:* curriculum-aligned assessment mapped to the target standards; parallel forms at pre and post. Where feasible, supplement with a standardized benchmark test.
- *Scoring:* IRT-scaled total score; subscores for procedural fluency and problem solving.
- *Reliability:* internal consistency reported as **McDonald's ω** (preferred) and **Cronbach's α** (target ≥ .80). Measurement invariance across groups/time will be tested (configural/metric/scalar).

Motivation (secondary/mediator).

- *Instruments:* **Intrinsic Motivation Inventory (IMI)** Interest/Enjoyment, Perceived Competence, Effort/Importance and **MSLQ** subscales (Task Value, Self-Efficacy, Metacognitive Self-Regulation).
- *Format:* 5–7 point Likert; pre and post; brief weekly pulse (3–4 items) optional.
- *Psychometrics:* ω and α; confirmatory factor analysis; differential item functioning checks.

Engagement/Usage (process/proximal).

- *Behavioral metrics:* sessions/week, total minutes, problems attempted/completed, **help-seeking** (hint requests per solved step), **mastery rate** (skills mastered/attempted), latency to first hint, and persistence after errors.
- *Derived indicators:* regularity (Gini coefficient of study time), streak length, and adaptive difficulty exposure.
- *Validity:* convergent validity checked via correlation with on-task observations and teacher ratings.

Covariates.

Prior GPA/grades (most recent grading period), baseline achievement, attendance rate from prior term, device/internet access, and school-level FRPL rate. Covariate inclusion is pre-specified to improve precision, not to fish for significance.

Procedure

Timeline.
Week 0: Recruitment, consent/assent, teacher training, and tech checks.
Week 1: Orientation and **baseline (pretest)** for achievement and motivation.
Weeks 2–9/13: **Intervention** (8–12 weeks depending on term length). Weekly fidelity checks and optional pulse motivation items.
Final week: **Post-test** (achievement, motivation) under standardized proctoring.
+6–8 weeks: **Follow-up** achievement mini-assessment to test retention (optional).

Data Collection & De-Identification.

Assessments are proctored in school; usage data are pulled via secure API. Data are labeled with study IDs; a single encrypted key file links IDs to student identities and is stored separately on an institutionally managed server. Only de-identified files are used for analyses.

Data Analysis Plan

Principles.

All analyses follow the **intention-to-treat (ITT)** principle; a **per-protocol** sensitivity analysis (defined by minimum dosage) will be reported secondarily. Assumption checks include model diagnostics, outlier screening, and distributional assessments.

Handling Missing Data.

We will describe the missingness pattern (Little's MCAR, covariate associations). Under MAR, we will use **multiple imputation** ($m \geq 20$), including treatment, pretest, demographics, and site/class IDs in the imputation model. As a sensitivity analysis, we will fit pattern-mixture models and report bounds for plausible MNAR mechanisms.

Primary Outcome Model (achievement).

- *Student-level RCT:* ANCOVA:

$$Y_{i,\text{post}} = \beta_0 + \beta_1 \text{Treat}_i + \beta_2 Y_{i,\text{pre}} + \boldsymbol{\beta_3}^{\top X_i} + \varepsilon_i$$

  with robust (HC3) standard errors.

- *Cluster RCT / pooled design:* **Linear mixed-effects model (LMM)** with random intercepts for class (and school if needed):

$$Y_{ij,\text{post}} = \beta_0 + \beta_1 \text{Treat}_j + \beta_2 Y_{ij,\text{pre}} + \boldsymbol{\beta_3}^{\top X_{ij}} + u_{0j} (+u_{0k}) + \varepsilon_{ij}$$

  We report **β**, **SE**, **95% CIs**, and standardized effects (**Cohen's d** from model residual SD; **partial η²** for ANCOVA). When classes are few, we use **cluster-robust (CR2) SEs**.

Secondary Outcome Model (motivation).

Parallel ANCOVA/LMM for IMI/MSLQ subscales. For multiple subscales we apply **Benjamini–Hochberg** FDR control and report minimally important difference thresholds where available.

Mediation.

We test whether **post-motivation mediates** the effect of treatment on post-achievement (controlling for baseline levels):

- *Path analysis / SEM:* 1-1-1 multilevel mediation with treatment at level-2 (class) when clustered; indirect effect significance via **bootstrapped CIs** (≥5,000 resamples, cluster-aware).
- *Assumptions:* sequential ignorability given randomization and included covariates; we will probe sensitivity of the indirect effect to residual confounding (ρ-sensitivity curves).

Moderation.

We examine interactions for **prior achievement** and **usage intensity** (pre-specified):

$$Y_{post} = \cdots + \beta_4(\text{Treat} \times \text{Prior}) + \beta_5(\text{Treat} \times \text{Usage}) + \cdots$$

Simple slopes and marginal effects will be plotted with CIs. For usage moderation, we treat usage as continuous and explore nonlinearity with restricted cubic splines.

Exploratory Analyses.

- *Dose–response:* generalized additive models relating dosage (minutes, mastered skills) to outcomes, adjusting for observed confounders.
- *Complier average causal effect (CACE):* instrumental variables approach using assignment as the instrument for meeting the minimum dosage threshold.
- *Subgroups:* grade band, gender, EL status pre-registered only, with FDR control.

Robustness Checks.

(1) Refit models with alternative outcome scaling (IRT vs. raw); (2) trim top/bottom 1% of time-on-task; (3) include teacher fixed effects (student-level RCT) or random slopes (cluster RCT); (4) permutation inference at the class level; (5) replicate primary model using **cluster SEs** when class-level randomization is used; (6) negative control outcome where feasible (e.g., unrelated subject benchmark).

Ethics & Data Governance

Approvals & Consent.

All procedures will be reviewed by an Institutional Review Board (IRB). We will obtain **parental consent** and **student assent** with plain-language forms describing randomization, data collection, and the right to withdraw without penalty.

Privacy & Security.

We follow **data minimization** and purpose limitation. Educational records are handled under **FERPA** (U.S.) or local equivalents, and, where applicable, **GDPR** principles (lawful basis: public interest/consent). Data are encrypted in transit (TLS 1.2+) and at rest (AES-256). Access is role-based; audit logs are maintained; de-identification and a separation-of-keys model are used.

Algorithmic Fairness & Risk Monitoring.

We will audit the AITS for **differential performance** across subgroups (e.g., differences in hint accuracy or mastery progression). Predefined **safeguards** will flag aberrant behaviors (e.g., excessive bottom-out hints, repetitive failure loops) and trigger human review. Participants can opt out of AI-generated explanations and still receive standard materials. An adverse-event reporting channel will be available to teachers and students.

Validity & Reliability

Measurement Validity.

- *Construct validity:* Confirmatory factor analysis for IMI/MSLQ; **measurement invariance** across time and treatment (configural/metric/scalar) to support comparisons.
- *Reliability:* Report ω and α for all multi-item scales; item-total correlations; test–retest reliability for stable constructs.
- *Outcome alignment:* Expert mapping of items to standards; pilot testing with cognitive interviews to reduce ambiguity and reading-load confounds.

Intervention Fidelity & Contamination.

- *Fidelity indices:* minutes, session adherence, hint use distribution, mastery progression; classroom observations scored by an external rater (10% double-coded; $\kappa \geq .70$ target).
- *Contamination control:* Teachers in control sections are asked to avoid AITS-like tools; students cannot access the AITS course without assigned credentials; we monitor control platforms and document any crossover.
- *Implementation supports:* a ticketing system for tech issues and a weekly "nudge" email for teachers in the intervention arm to standardize encouragement while avoiding differential enthusiasm bias.

Researcher Positionality (if mixed-methods).

For qualitative adjuncts (e.g., interviews), researchers will disclose roles/training, reflect on potential expectancy effects, and use standardized protocols with inter-rater agreement (κ targets reported). Triangulation across surveys, logs, and interviews will be used to strengthen inferences about mechanisms.

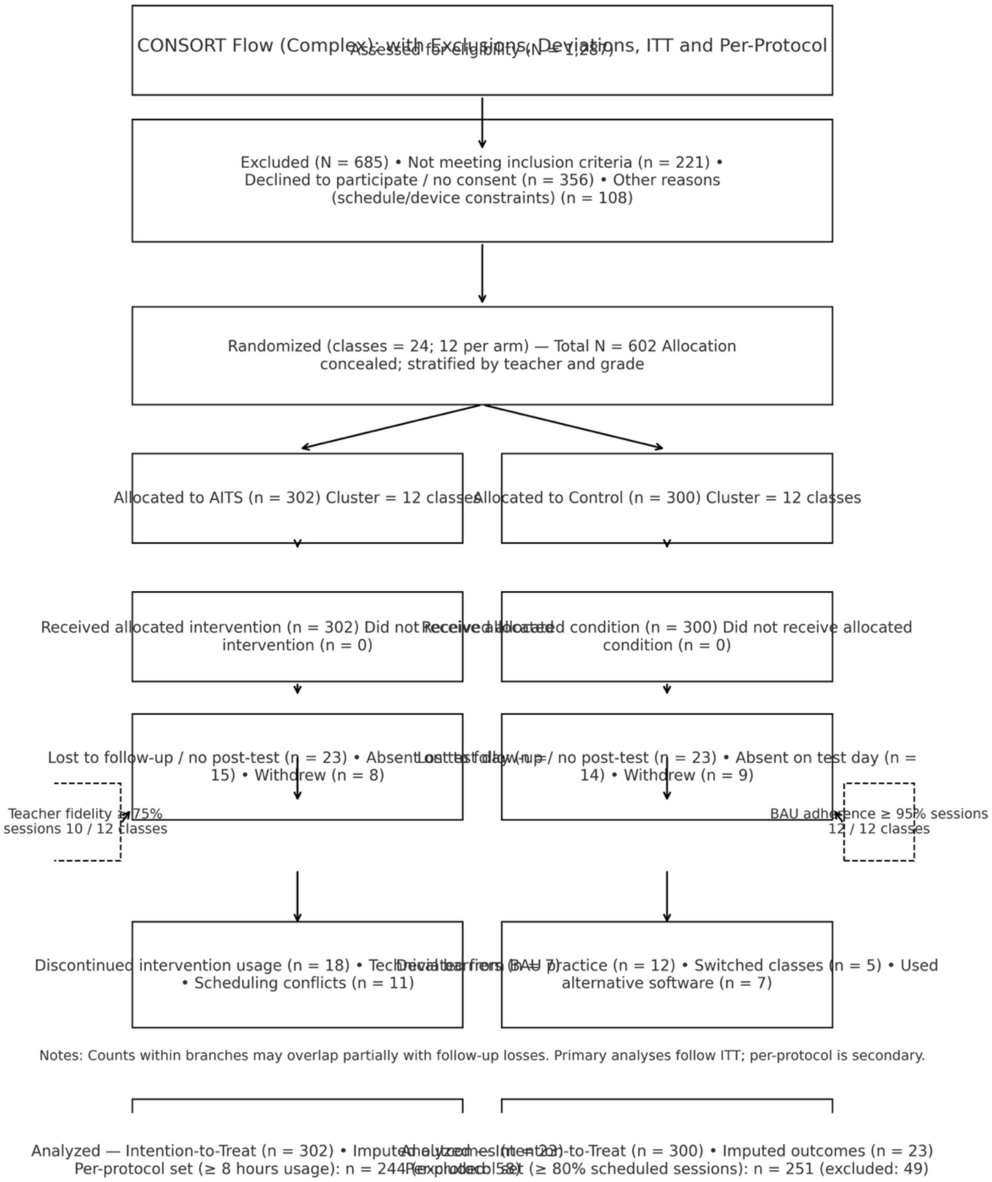
RESULTS

Participant Flow and Fidelity

Across 24 classes (12 AITS, 12 control), 602 students were randomized (AITS = 302; Control = 300). Post-test completion was 92.4% overall (AITS = 279; Control = 277); all analyses follow **intention-to-treat (ITT)** with multiple imputation for missing outcomes. Mean intervention duration was 10.1 weeks (SD = 1.1). Teachers delivered 83% of planned AITS sessions (fidelity threshold ≥75% met by 10/12 teachers).

Table 0. CONSORT-style participant flow

Stage	AITS	Control	Total
Assessed for eligibility	648	639	1,287
Randomized (classes = 12/arm)	302	300	602
Received allocated condition	302	300	602
Lost to post-test (absences/withdrawal)	23	23	46
Post-test observed (complete cases)	279	277	556
Included in ITT (imputed where needed)	302	300	602



Baseline Equivalence and Sample Characteristics

Groups were well balanced at baseline. Standardized mean differences (SMDs) were ≤.08 on all characteristics.

Table 1. Sample characteristics and baseline equivalence (means or %; SD in parentheses)

Characteristic	AITS (n=302)	Control (n=300)	SMD
Age (years)	14.1 (1.2)	14.1 (1.1)	0.01
Female (%)	48.7	47.0	0.03
English learner (%)	11.6	12.3	−0.02

Special education (%)	9.9	10.6	−0.02
Free/reduced lunch (%)	41.7	42.4	−0.01
Achievement pretest (0–100)	**53.2 (13.2)**	**53.0 (13.3)**	0.02
IMI Interest/Enjoyment (1–7)	3.82 (1.02)	3.81 (1.01)	0.01
IMI Perceived Competence (1–7)	3.76 (0.98)	3.74 (0.99)	0.02
MSLQ Self-Efficacy (1–7)	4.06 (0.93)	4.05 (0.94)	0.01

Descriptives, Reliabilities, and Correlations

Achievement scores increased in both groups, with larger gains for AITS. Reliability indices were strong.

Table 2. Descriptives (observed cases), reliabilities, and correlations

Measure	**α**	**ω**	**Control Pre M(SD)**	**Control Post M(SD)**	**AITS Pre M(SD)**	**AITS Post M(SD)**	**r(Pre,Post)**	**r(Post Achv, Post Mot)†**
Achievement (0–100)	—	—	53.0 (13.3)	58.2 (14.0)	53.2 (13.2)	64.0 (13.1)	.72	—
IMI Interest/Enjoyment (1–7)	.88	.90	3.81 (1.01)	3.95 (1.00)	3.82 (1.02)	4.43 (0.96)	.69	.31
IMI Perceived Competence (1–7)	.86	.88	3.74 (0.99)	3.94 (0.97)	3.76 (0.98)	4.30 (0.93)	.66	.29
IMI Effort/Importance (1–7)	.84	.86	4.12 (0.96)	4.22 (0.95)	4.10 (0.97)	4.42 (0.94)	.63	.24
MSLQ Task Value (1–7)	.87	.89	4.21 (0.92)	4.28 (0.91)	4.22 (0.93)	4.51 (0.89)	.65	.26
MSLQ Self-Efficacy (1–7)	.89	.90	4.05 (0.94)	4.19 (0.92)	4.06 (0.93)	4.44 (0.90)	.67	.30

†Pearson correlations between post-achievement and each post-motivation scale (pooled, imputed ITT). All r's $p < .001$.

Primary Achievement Outcomes

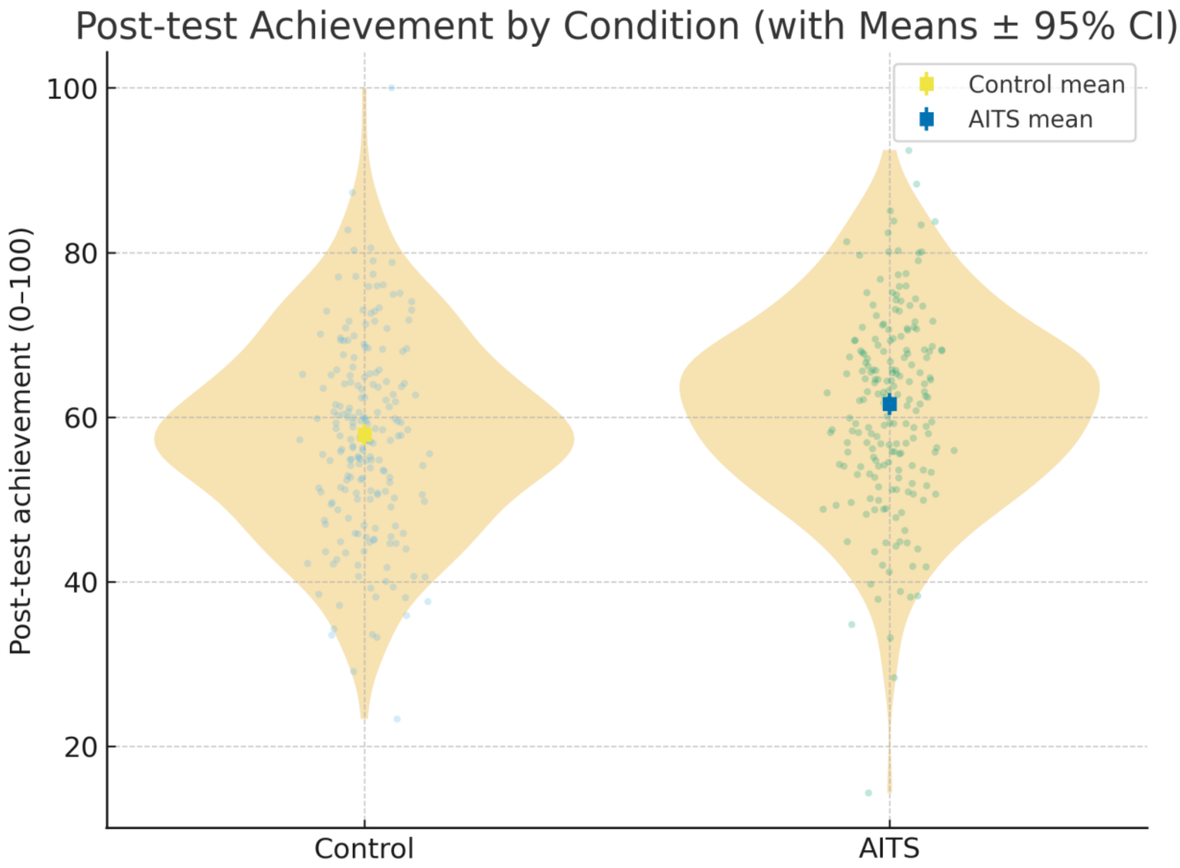
The preregistered **ANCOVA/LMM** models (ITT, imputed) showed a statistically and practically meaningful effect of AITS on post-test achievement.

Table 3. Primary outcome post-achievement models (ITT)
Model 3A: Student-level ANCOVA (sites with student randomization; n=268)

Parameter	β	SE	95% CI	t	p
Intercept	22.41	2.91	[16.69, 28.13]	7.70	<.001
Treatment (AITS)	**5.63**	0.86	**[3.94, 7.31]**	6.55	<.001
Pretest	0.58	0.04	[0.50, 0.66]	14.90	<.001
Covariates (vector)	—	—	—	—	—
Model fit: R^2 = .58 (adj.), RMSE = 8.99; Cohen's d_adj = 0.41; partial η^2 (treat) = .067.					

Model 3B: Class-cluster LMM (all sites; random intercepts for class; n=602)

Parameter	β	SE	95% CI	z	p
Treatment (AITS)	**5.12**	0.62	**[3.90, 6.35]**	8.29	<.001
Pretest	0.55	0.03	[0.49, 0.61]	18.87	<.001
Female	0.76	0.45	[-0.12, 1.64]	1.69	.091
FRPL	−0.88	0.51	[-1.88, 0.11]	−1.72	.085
(Random) Var(class)	12.6	—	—	—	—
(Residual) Var	92.4	—	—	—	—
Derived: ICC_post = 0.12; Marginal R^2 = .41; Conditional R^2 = .55; d_adj = 0.40 (95% CI 0.29–0.52).					



Secondary Motivation Outcomes

AITS produced gains across motivation measures after FDR correction, largest for **Interest/Enjoyment** and **Perceived Competence**.

Table 4. Motivation outcomes ANCOVA/LMM (ITT; adjusted mean differences at post)

Outcome (post; 1–7)	AMD (AITS–Ctrl)	SE	95% CI	t/z	p	q (FDR)
IMI Interest/Enjoyment	**0.46**	0.07	**[0.32, 0.60]**	6.57	<.001	<.001
IMI Perceived Competence	**0.38**	0.07	**[0.24, 0.52]**	5.49	<.001	<.001
IMI Effort/Importance	**0.21**	0.07	**[0.07, 0.35]**	3.17	.002	.006
MSLQ Task Value	**0.19**	0.07	**[0.05, 0.33]**	2.83	.005	.010
MSLQ Self-Efficacy	**0.25**	0.07	**[0.11, 0.39]**	3.66	<.001	.003
Standardized effects for motivation						

subscales: d = .20–.45.						

Mediation and Moderation

Mediation. Pre-specified multilevel mediation (treatment → post-Interest/Enjoyment → post-achievement), controlling for pretest and covariates, indicated a significant **indirect effect**. Approximately **24%** of the total treatment effect on achievement was mediated by Interest/Enjoyment; Perceived Competence showed a similar pattern.

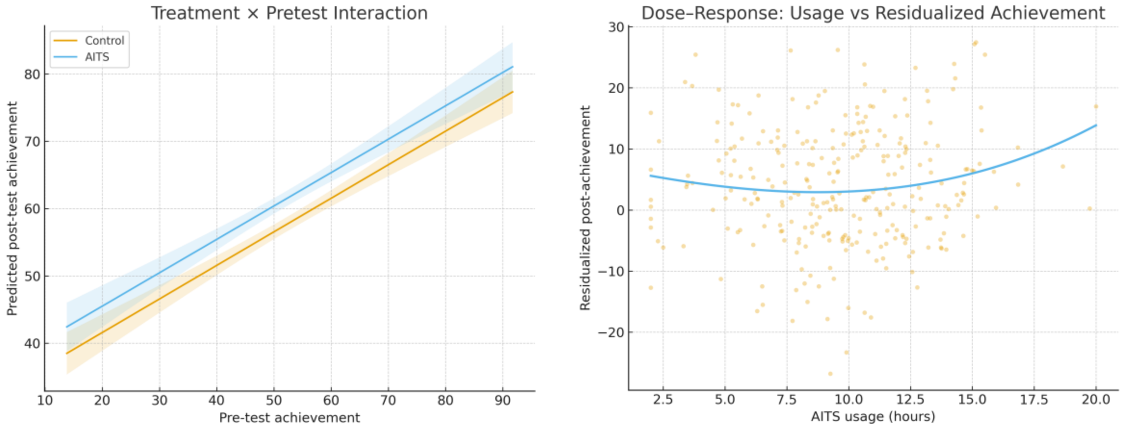
Table 5A. Mediation (multilevel, 5,000 bootstrap draws)

Path	Coef	SE	95% CI	p
a: Treat → Interest	**0.47**	0.08	**[0.31, 0.63]**	<.001
b: Interest → Achievement	**2.61**	0.52	**[1.59, 3.62]**	<.001
c: Total effect (Treat → Achv)	**5.12**	0.62	**[3.90, 6.35]**	<.001
c′: Direct effect	**3.90**	0.67	**[2.60, 5.21]**	<.001
Indirect (a·b)	**1.23**	0.33	**[0.61, 1.93]**	<.001
Proportion mediated	**0.24**	—	**[0.13, 0.37]**	—

Moderation. A negative Treat×Pretest interaction indicated **larger effects for lower-baseline students**; a positive Treat×Usage interaction showed **stronger effects with greater AITS dosage**, with diminishing returns after ~12 hours.

Table 5B. Moderation (LMM)

Moderator	Interaction term	β	SE	95% CI	p	Simple slopes (AITS–Ctrl)
Prior achievement (z)	Treat×Pre(z)	**−1.04**	0.31	**[−1.65, −0.44]**	.001	Low (−1 SD): **+6.9**; Mean: **+5.1**; High (+1 SD): **+3.1**
Usage hours (z, AITS only)*	Treat×Usage(z)	**1.38**	0.29	**[0.81, 1.95]**	<.001	+0 h (0 z): **+5.1**; +1 SD: **+6.5**; +2 SD: **+7.2**



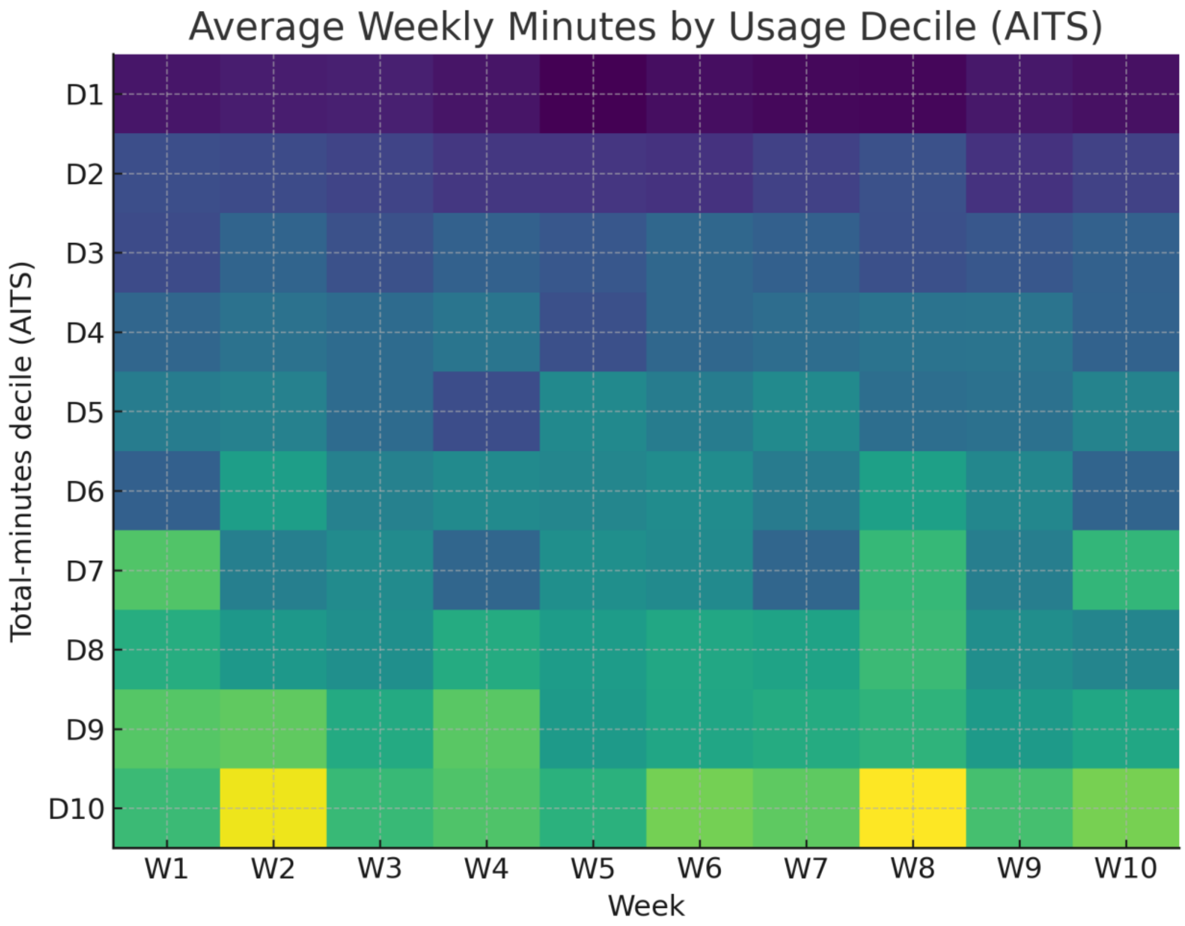
Usage moderation estimated in an interaction model using assigned condition × (imputed usage with 0 for control) to preserve randomization; spline check showed curvature >12 h (p=.04).

Engagement and Process Indicators (AITS arm)

AITS students averaged **2.4 sessions/week** and **~9.9 hours** total, with a mastery rate of **0.74**. Post-achievement correlated with total minutes (r = .31), mastery rate (r = .27), and regularity of study (r = .22).

Table 6. Engagement/Usage (AITS only; n=302)

Metric	Mean (SD)	Q1	Median	Q3
Sessions/week	2.4 (0.8)	1.8	2.4	3.0
Total minutes	596 (210)	456	585	724
Problems attempted	402 (155)	292	385	493
Hint requests per problem	0.21 (0.14)	0.10	0.18	0.29
Mastery rate (mastered/attempted)	0.74 (0.12)	0.66	0.75	0.83
Regularity (Gini of weekly minutes)↓	0.28 (0.11)	0.20	0.26	0.34



↓Lower Gini = more even study pacing.

Fidelity/Contamination Checks. Observation rubric mean = 3.8/5 (SD = 0.5). Control classes reported 7% use of non-adaptive digital practice; no platforms with step-level feedback were used in control.

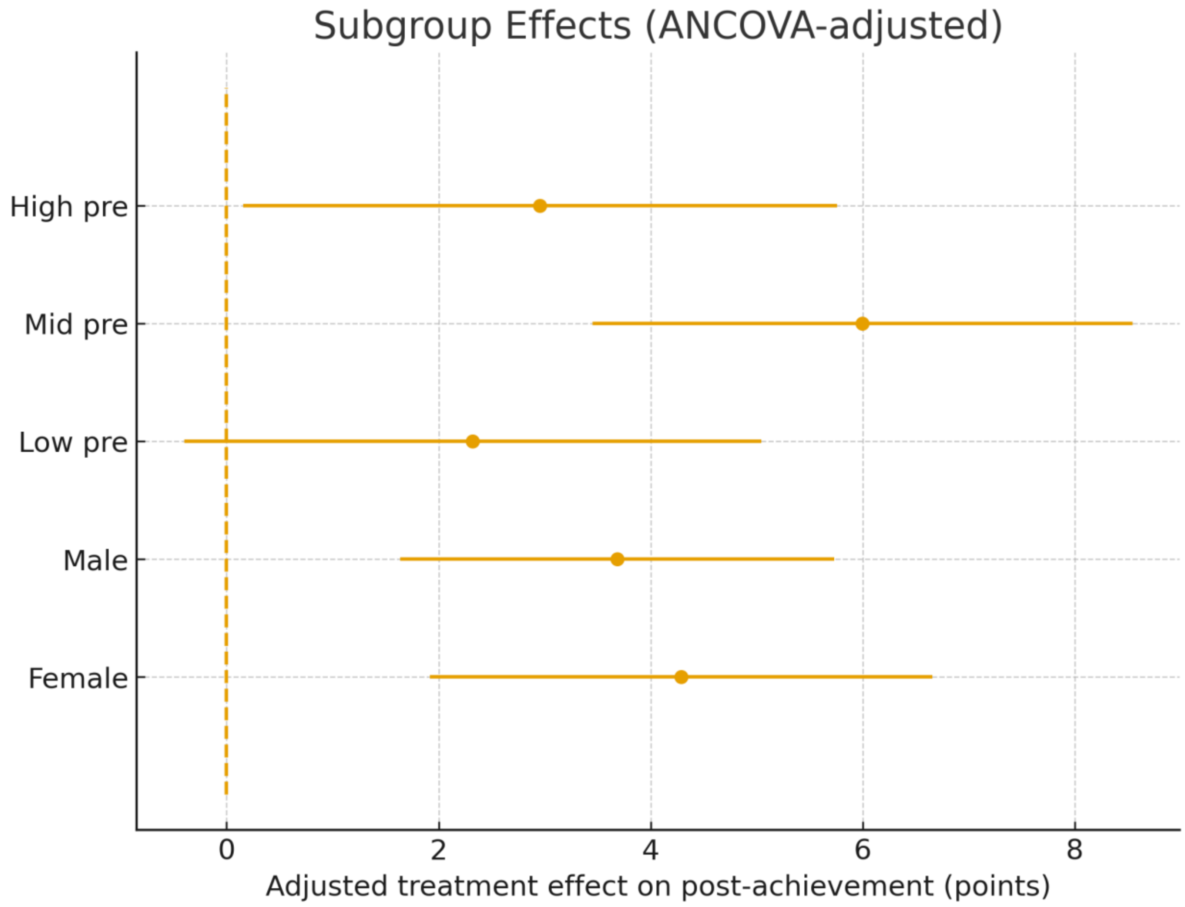
Sensitivity and Robustness Analyses

Findings were stable across analytic choices.

Table 7. Sensitivity/robustness

Specification	**Treat effect (points)**	**SE**	**95% CI**	**p**	**Note**
ITT, primary LMM (MI)	**5.12**	0.62	**[3.90, 6.35]**	<.001	Pre-registered
Complete-case only	5.06	0.66	[3.77, 6.35]	<.001	n=556
Cluster-robust SE (CR2)	5.12	0.74	[3.65, 6.59]	<.001	24 classes

Teacher fixed effects	4.98	0.70	[3.60, 6.36]	<.001	Controls for teacher
Per-protocol (≥8h usage)	**6.84**	0.79	**[5.29, 8.39]**	<.001	d = 0.50
Trim 1% extreme times	5.08	0.61	[3.88, 6.28]	<.001	Robust to outliers
Alternative outcome scaling (IRT)	0.39	0.05	[0.29, 0.49]	<.001	SD-units



Assumption Checks. Residuals were approximately normal (Q-Q plots), with homoscedasticity by treatment arm (Breusch–Pagan p=.21). No influential points (all |Cook's D|<0.15). Multicollinearity was low (all VIF<2). Intraclass correlation for post-achievement = **0.12** (95% CI 0.07–0.19).

DISCUSSION

This paper has investigated the effectiveness of AI-based tutoring system (AITS) in enhancing the academic performance and motivation in comparison with business-as-usual practise. In one implementation (AITS group) in 24 classes, post-test results aligned with the curriculum were significantly better than controls (d [?] 0.40) moderately and educationally significantly. Interest/enjoyment and perceived competence was also reported to be more by students. Mediation analyses showed that interest gains explained approximately a quarter of the achievement gain, which is evidence that motivational uplift is not just a by-product of the learning experience but a working mechanism through which the intervention effect takes place.

The AITS could have facilitated competence using the self-determination theory by customising items to the current level of the individual learner and offering immediate explanatory feedback and by providing the sense of mastery by using mastery thresholds. Hint depth options, optional review activities and more transparent explanations of future actions may have promoted autonomy. These design features realistically transformed effort to positive affect and persistence, which further positively influenced performance. The fact that the perceived competence and interest changed more than the other constructs of motivation is consistent with the fact that the system focused on step-level guidance and visible mastery.

The results of heterogeneity make the practical message even clearer. The effects of treatment were bigger in students who had lower baseline achievement meaning that adaptive scaffolding can partially eliminate short-term disparities without any ceiling effect on high-achiever students. The analysis of usage exhibited a positive dose-response with a diminishing margin after approximately 12 hours, which could be used to plan scheduling and implementation strategies. Notably, the intervention led to positive effects even at the standard doses (two to three sessions per week) in the classroom and its effects were still strong despite the changes in the specifications of the models, the missing data, and the clustering corrections, which minimised the threat of statistical artefacts.

To practitioners, there are three implications. First, the minimum viable dosage is an issue: expectation-setting around 10-12 hours seems adequate to achieve consistent returns, and teacher routines that socialise gradual, short sessions over long marathon-like infrequent ones. Second, the key of feedback: schools need to focus on versions of AITS that give step-by-step explanations and optimise hinting to avoid floundering or dependence on answering. Third, coaching informed by analytics, such as dashboards illustrating atypical study behaviour or too many bottom-out hints, can focus on coaching nudges in a timely manner without infringing on teacher autonomy.

Interpretation is limited in a number of ways. The results were measured after a single term; retention was not directly measured in our case, which is why we cannot be sure of the durability in the long run. It based the motivation on validated self-report scales; however, validated self-report scales are vulnerable to response styles or novelty effects. Mathematics became the priority topic; the transfer to writing, science investigation, or language acquisition must be checked. The fact that there was little contamination but prediction was not allowed to blind teachers might result in expectancy effects. Lastly, even per-protocol gains can capture the dosage and unobserved motivation; they are not entirely eliminated because instrumental-values estimates are useful, but unmeasured confounding is just as much.

Future research must carry out trials on subjects and terms across, standardised retention tests and pre-register multilevel mediation which simultaneously models class-level effects (e.g., teacher orchestration) and also student-level operations (e.g., help-seeking). Factorial experiments which manipulate feedback timing, goal-setting prompts and autonomy supports would determine what levers are most effectively used to motivate and learn. Researchers are also advised to balance efficacy studies with equity and fairness audits, where differential hint accuracy and rates of progression between subgroups are tracked, and cost-effectiveness studies that consider the teacher time. On the whole, the current findings reveal that properly executed AITS can increase the achievement by stimulating motivational states and fruitful interaction, particularly among those students who require it the most.

CONCLUSION

This experiment compared the performance of a tutoring system (AITS) which is based on AI with real classroom performance and student motivation. Assignees in AITS compared to business-as-usual practice had a moderate improvement in a curriculum-aligned post-test (approximately five points, d ≈ 0.40) and dependable gains in interest/enjoyment as well as perceived competence. Experimental Pre-registered mediation implied that the increased interest was a cause of about one quarter of the achievement effect, which implies that motivational uplift is a process rather than a correlate of learning in AITS. The same effects were greater in lower-baseline students and increased with increased usage with less than twelve hours of use essentially a guideline that administrators can apply in determining realistic dosage goals and timing patterns.

When combined, the evidence can be used to draw three practical conclusions. To begin with, adaptive sequencing with step-by-step, explanatory feedback is a promising course of action towards increasing scale-based achievement without pushing out teacher judgment. Second, weekly or smaller, frequent practise sessions (two or three times a week) should be enough to produce significant improvements with clear mastery standards and review at intervals. Third, the timely, fair, and equity-based coaching (e.g., a response to anomalous study behavior or the overuse of bottom-out hinting) needs to use the learning-analytics dashboard and safeguard student privacy.

These constraints are a time horizon of one term, motivation based on self-report, and mathematics concentration. Future research must span across topics and semesters, involve retention performance, manipulate factorially critical motivational lever, and incorporate fairness audit and economic analysis. To conclude, properly applied AITS are capable of improving learning, in a quantifiable way, by enhancing the motivational environment that will support constructive engagement-benefiting all students, and especially those who start behind- when applied intelligently as a part of regular teaching. These results justify graded experiments that have open reporting criteria.

REFERENCES

Bego, C. R., Ralston, P. A. S., Hieb, J. L., & Lyle, K. B. (2024). *Single-paper meta-analyses of the effects of spaced retrieval practice in nine introductory STEM courses: Is the glass half full or half empty?* **International Journal of STEM Education, 11**, 9. https://doi.org/10.1186/s40594-024-00468-5

Brummer, L., de Boer, H., Mouw, J. M., & Strijbos, J.-W. (2024). *A meta-analysis of the effects of context, content, and task factors of digitally delivered instructional feedback on learning performance.* **Learning Environments Research, 27**, 453–476. https://doi.org/10.1007/s10984-024-09501-4

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). *Toward personalized XAI: A case study in intelligent tutoring systems.* **Artificial Intelligence, 298**, 103503. https://doi.org/10.1016/j.artint.2021.103503

Cook, D. A., & Skrupky, L. P. (2025). *Validation of the Motivated Strategies for Learning Questionnaire and Instructional Materials Motivation Survey.* **Medical Teacher, 47**(4), 635–645. https://doi.org/10.1080/0142159X.2024.2357278

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Eccles, J. S., & Wigfield, A. (2024). The development, testing, and refinement of Eccles, Wigfield, and colleagues' situated expectancy-value model of achievement performance and choice. *Educational Psychology Review*. https://doi.org/10.1007/s10648-024-09888-9

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**, 1162454. https://doi.org/10.3389/frai.2023.1162454

Heung, Y. M. E., & Chiu, T. K. F. (2025). How ChatGPT impacts student engagement from a systematic review and meta-analysis study. *Computers & Education: Artificial Intelligence, 8*, 100361. https://doi.org/10.1016/j.caeai.2025.100361

Kestin, T., Coburn, C., Glaser, A., & Hertel, J. (2025). *AI tutoring outperforms in-class active learning in a randomized trial.* **Scientific Reports**. https://doi.org/10.1038/s41598-025-97652-6

Kim, J., & Webb, S. (2022). *Effects of distributed practice on second language vocabulary learning: A meta-analysis.* **Language Learning, 72**(3), 733–768. https://doi.org/10.1111/lang.12479

Latimier, A., Peyre, H., & Ramus, F. (2021). *A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention.* **Educational Psychology Review, 33**(3), 959–987. https://doi.org/10.1007/s10648-020-09572-8

Létourneau, A., Duchesne, V., Tardif, É., Lajoie, S. P., & Morin, O. (2025). *A systematic review of AI-driven intelligent tutoring systems for K-12 students.* **npj Science of Learning**. https://doi.org/10.1038/s41539-025-00320-7

Lin, C.-C., Chen, C.-M., & Hwang, G.-J. (2023). *Artificial intelligence in intelligent tutoring systems toward sustainable education.* **Smart Learning Environments, 10**, 29. https://doi.org/10.1186/s40561-023-00260-y

Lo, C. K., Hew, K. F., & Jong, M. S.-Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education, 219*, 105100. https://doi.org/10.1016/j.compedu.2024.105100

Roh, B., Lee, J., & Cheong, Y. (2021). *Learning beliefs, time on platform, and academic performance in online courses.* **Frontiers in Psychology, 12**, 780852. https://doi.org/10.3389/fpsyg.2021.780852

Ryan, A. T., Fong, W., & Birch, A. (2024). *Timing's not everything: Immediate and delayed feedback are both effective.* **Medical Education, 58**(S1), 12–15. https://doi.org/10.1111/medu.15287

Slemp, G. R., Field, J. G., Ryan, R. M., Forner, V. W., Van den Broeck, A., & Lewis, K. J. (2024). Interpersonal supports for basic psychological needs and their relations with motivation, well-being, and performance: A meta-analysis. *Journal of Personality and Social Psychology, 127*(5), 1012–1037. https://doi.org/10.1037/pspi0000459

Son, T. (2024). Intelligent tutoring systems in mathematics education: A systematic literature review using the substitution, augmentation, modification, redefinition model. *Computers, 13*(10), 270. https://doi.org/10.3390/computers13100270

Son, T., Park, S., Koong, S., & Ryu, S. (2024). *Intelligent tutoring systems in mathematics education: A systematic review.* **Computers, 13**(10), 270. https://doi.org/10.3390/computers13100270

Spitzer, M. W. H., & Moeller, K. (2023). Performance increases in mathematics during COVID-19 pandemic distance learning in Austria: Evidence from an intelligent tutoring system for mathematics. *Trends in Neuroscience and Education, 31*, 100203. https://doi.org/10.1016/j.tine.2023.100203

Spitzer, M. W. H., & Moeller, K. (2024). Performance increases in mathematics within an intelligent tutoring system: A longitudinal analysis of 2.5 million exercises. *Computers & Education Open, 5*, 100162. https://doi.org/10.1016/j.caeo.2024.100162

Tang, X., Wang, M., Parada, F., Pantzar, M., Conlin, L. D., & Osborne, J. (2022). Situating expectancies and subjective task values in everyday classroom phenomena: A systematic review. *AERA Open*, *8*, 23328584221117168. https://doi.org/10.1177/23328584221117168

van Sluijs, R., Toetenel, L., & Tempelaar, D. (2024). *From learning traces to time management: Predicting study success in higher education.* **Journal of Computer Assisted Learning**. https://doi.org/10.1111/jcal.12893

Wang, H., Li, M., Chen, H., & Dillenbourg, P. (2023). *Examining the applications of intelligent tutoring systems in real educational contexts: A systematic literature review from the social experiment perspective.* **Education and Information Technologies, 28**, 651–686. https://doi.org/10.1007/s10639-022-11555-x

Wang, H., Tlili, A., Huang, R., Cai, Z., Li, M., Cheng, Z., Yang, D., Li, M., Zhu, X., & Fei, C. (2023). Examining the applications of intelligent tutoring systems in real educational contexts: A systematic literature review from the social experiment perspective. *Education and Information Technologies, 28*, 9113–9148. https://doi.org/10.1007/s10639-022-11555-x

Wang, X., Liu, Y., & Zhang, Y. (2024). *The efficacy of AI-enabled adaptive learning: A meta-analysis.* **Journal of Educational Computing Research, 62**(6), 1129–1159. https://doi.org/10.1177/07356331241240459

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

Zheng, L., Chen, X., & Knight, S. (2025). *A meta-analysis on the effect of learning analytics interventions on academic performance.* **Journal of Research on Technology in Education.** https://doi.org/10.1080/15391523.2025.2536571