AI in the Classroom: Barrier or Gateway to Academic and Labor Market Success?

Catalina Franco* Natalie Irmert† Siri Isaksson‡

November 14, 2025

Abstract

Artificial Intelligence (AI) is becoming an increasingly important skill in the labor market, but will it also affect academic success? Recent research shows that current students – who will be facing this rapidly changing labor market – are adopting AI tools at differential rates based on both gender and ability. Whether AI will affect adopters' academic success hinges on whether AI interferes with or enhances learning, which in turn depends on whether AI is being used as a substitute for or complement of own effort. We run a controlled lab experiment with 478 students learning Esperanto across three treatments: control (no AI), AI-assisted (unguided AI access), and AI-guided (structured AI guidance). Our results show no significant main treatment effects on test performance. However, we find suggestive evidence of heterogeneous effects by gender and prior ability (GPA). The findings provide evidence on whether differential AI adoption by gender and ability is likely to create gaps in academic success, and explore mechanisms including complement/substitute behavior, confidence, cheating perceptions, and motivation.

*Center for Applied Research (SNF) at NHH – Norwegian School of Economics

†Lund University

‡Norwegian School of Economics

1 Introduction

The rapid emergence of generative AI tools, particularly ChatGPT, has created both opportunities and challenges for educational institutions worldwide. Recent research shows that AI is poised to reshape different parts of the economy, with emerging evidence that various factors determine early adoption of AI. Gender and ability have been identified as key determinants of AI use (1). For instance, top female students tend to opt out of AI use, raising concerns about potential widening of existing gender gaps in STEM fields and labor market outcomes.

At the same time, educational institutions are grappling with fundamental questions: Is AI beneficial or harmful for learning? Can institutions ensure that everyone reaps the benefits of AI by guiding student use? These questions are particularly urgent given that students entering the labor market will face increasing demands for AI proficiency.

This paper addresses these questions through a carefully designed laboratory experiment with 478 university students. We implement three between-subject treatment variations: (1) a control group with no ChatGPT access (but with Google Search, representing pre-AI learning), (2) an AI-assisted treatment with ChatGPT access but no formal guidance, and (3) an AI-guided treatment with ChatGPT access plus structured guidance on effective use. Students learn Esperanto – a language virtually unknown to participants – allowing us to measure genuine learning rather than pre-existing knowledge.

Our main findings reveal no significant average treatment effects of AI access on test performance. Mean test scores are remarkably similar across treatments: 8.13 in control, 8.18 in AI-assisted, and 7.99 in AI-guided groups. However, we find suggestive evidence of heterogeneous effects by gender and prior academic ability. While these interactions do not reach conventional significance levels, the patterns suggest that AI may affect different student subgroups differently.

We also document significant impacts on student effort during the learning phase. Students in AI treatments attempted substantially fewer practice questions compared to controls, consistent with AI serving as a substitute for traditional learning effort. This finding has important implications for understanding how AI tools reshape learning

behaviors.

Our mechanism analysis explores four key channels: (1) whether students use AI as a complement versus substitute for learning, (2) confidence and preparedness effects, (3) cheating perceptions, and (4) motivation and engagement. The results suggest complex interactions between AI access and student psychology that may not manifest immediately in test scores but could have longer-term implications.

This paper contributes to the growing literature on AI in education by providing experimental evidence on causal effects of AI access on learning outcomes. Unlike observational studies of AI adoption, our randomized design allows us to isolate the impact of AI from selection effects. The findings are directly relevant to policy debates about whether and how to integrate AI into formal curricula.

The remainder of the paper proceeds as follows. Section 2 reviews related literature. Section 3 describes the experimental design and data. Section 4 presents the main results on test performance. Section 5 examines heterogeneous effects by gender and GPA. Section 6 analyzes mechanisms. Section 7 discusses implications and concludes.

2 Related Literature

Our study connects to several strands of literature. First, we contribute to emerging research on AI adoption in education. **(author?)** (2) document rapid but uneven adoption of ChatGPT among students, with significant gender and ability gaps. **(author?)** (1) find that top female students are particularly likely to opt out of AI use, potentially putting them at a disadvantage in AI-intensive labor markets.

Second, our work relates to the literature on technology and learning. Previous research on calculator use, internet access, and educational software has shown mixed effects on student achievement (3). A key question is whether new technologies serve as complements or substitutes for traditional learning effort. Our finding that AI reduces practice question attempts suggests substitution may be occurring.

Third, we contribute to research on gender gaps in STEM and technology adoption.

Women have historically been underrepresented in computer science and related fields (4). If AI becomes a critical skill for future labor markets, differential adoption rates could exacerbate existing gender inequality.

Fourth, our mechanism analysis connects to educational psychology literature on metacognition, self-efficacy, and learning strategies (5). Understanding not just whether AI affects learning, but how and why, is crucial for designing effective interventions.

Finally, our experimental approach builds on a tradition of lab experiments in economics education, demonstrating the value of controlled settings for identifying causal effects while maintaining external validity through realistic learning tasks.

3 Experimental Design and Data

3.1 Experimental Procedures

We conducted experiments with 478 students at the University of Nottingham's CeDEx Laboratory between September and November 2024. Students were recruited from undergraduate programs across the university through standard lab recruitment procedures. Upon arrival, students were randomly assigned to one of three treatment conditions.

All sessions followed the same four-stage procedure:

Stage 1 (15 minutes): Students studied written learning materials about Esperanto, a constructed language that very few participants had prior exposure to. All students received identical materials regardless of treatment.

Stage 2 (20 minutes): Students completed practice questions with access to learning aids depending on treatment assignment. This is the stage where treatment variations were implemented:

- *Control (T1):* Access to Google Search and Google Translate, but AI sites blocked (mimicking pre-ChatGPT learning)
- *AI-Assisted (T2):* Access to ChatGPT through a provided account with cleared memory, plus Google Search

- *AI-Guided (T3):* Same as T2, plus written guidance from researchers on effective ChatGPT use for learning

Stage 3: Students completed a 15-question test without access to any learning aids or notes. Test performance is our primary outcome variable.

Stage 4: Students completed a post-study survey measuring demographics, learning experience, and attitudes toward AI.

3.2 Incentive Structure

Students received:

- GBP 5 for completing the study
- GBP 7 for correctly answering at least 20 practice questions
- GBP 1 per correct answer on the test (maximum GBP 15)

The maximum possible payment was GBP 27. Students were repeatedly informed that the test was the most important and lucrative component, incentivizing them to focus on genuine learning rather than merely completing practice questions.

3.3 Sample and Balance

Table 1 presents summary statistics and balance tests. Our final analysis sample includes 478 students distributed as: Control (n=160), AI-Assisted (n=165), AI-Guided (n=153). We exclude observations from a small pilot study and students who violated test protocols by leaving the browser during the exam.

The randomization achieved good balance on observable characteristics. Female representation ranges from 44-48% across treatments, with no significant differences. High GPA students (first-class honours, 70%+) comprise 58-62% of each treatment group. Mean age is approximately 21 years across all treatments.

3.4 Key Variables

Primary Outcome: Test score (0-15), measuring number of correct answers on the Stage 3 test taken without learning aids.

Secondary Outcomes:

- Top scorer: Indicator for scoring above 10 correct answers
- Low scorer: Indicator for scoring below 5 correct answers
- Practice questions attempted: Number of questions completed during Stage 2
- Practice questions correct: Number of correct answers during practice

Heterogeneity Variables:

- Female: Gender indicator
- High GPA: Indicator for first-class honours (70%+)

Mechanism Variables: We construct four summary indices from post-survey questions measuring:

1. Complement vs. Substitute: Whether AI enhanced or replaced learning effort
2. Confidence: Comfort with tools and preparedness for test
3. Cheating Perceptions: Whether students viewed AI use as academic dishonesty
4. Motivation: Engagement and effort during learning

Each index averages responses across multiple related survey items, with appropriate reverse-coding of negatively-worded questions.

4 Main Results: Treatment Effects on Learning

4.1 Test Performance

Figure 1 presents our main results showing test performance across treatments. The average test scores are remarkably similar: 8.13 (SE=0.23) in Control, 8.18 (SE=0.23)

in AI-Assisted, and 7.99 (SE=0.24) in AI-Guided. Neither AI treatment produces a statistically significant difference from the control group.

Table 2 presents regression results formally testing these differences. Column (1) shows the baseline specification:

$$\text{TestScore}_i = \alpha_0 + \alpha_1 \text{AI-Assisted}_i + \alpha_2 \text{AI-Guided}_i + \varepsilon_i \tag{1}$$

The coefficient on AI-Assisted is 0.05 (SE=0.33, p=0.88), and on AI-Guided is -0.14 (SE=0.34, p=0.68). Neither coefficient is statistically or economically significant. The 95% confidence intervals are tight enough to rule out effects larger than approximately 0.7 points (less than half a standard deviation).

Column (2) adds demographic controls (gender, age, GPA) with virtually no change to the treatment coefficients. Column (3) adds session fixed effects to account for any time-of-day or day-of-week variation, again with minimal impact. The robustness of the null result across specifications strengthens our conclusion that AI access alone does not significantly improve or harm test performance on average.

4.2 Distribution of Scores

Figure 1, Panel C shows the full distribution of test scores by treatment. The distributions are remarkably similar, with most students scoring between 6-10 correct answers. We see no evidence that AI access shifts the entire distribution or affects only the tails.

We formally test whether AI affects the likelihood of extreme performance. Table 3 presents results for top scorers (¿10 correct) and low scorers (¡5 correct). The proportion of top scorers is 13.8% in Control, 13.9% in AI-Assisted, and 11.1% in AI-Guided – differences that are not statistically significant. Similarly, low scorer rates are 14.4%, 15.8%, and 17.0% respectively, again with no significant differences.

4.3 Practice Phase Behavior

While AI does not affect test scores, it significantly affects behavior during the learning phase. Figure 4, Panel A shows that students in AI treatments attempted substantially fewer practice questions. Control students averaged 43.2 practice questions, compared to 36.8 in AI-Assisted (p¡0.01) and 35.4 in AI-Guided (p¡0.01).

This 15-18% reduction in practice effort is consistent with AI serving as a substitute for traditional learning methods. Students with AI access may rely on the tool to provide quick answers rather than working through problems independently.

Interestingly, this reduction in quantity does not translate to lower quality. Figure 4, Panel C shows that accuracy rates on practice questions are similar across treatments (approximately 75-80%). This suggests students with AI access are being more selective about which questions to attempt rather than attempting more questions with lower accuracy.

4.4 Power and Equivalence Tests

Figure 9 presents comprehensive statistical evidence confirming our null findings. Panel B shows power analysis indicating we had 80% power to detect effects as small as Cohen's d = 0.35, which is considered a small-to-medium effect size. Our null result is therefore not due to insufficient sample size.

Panel D presents equivalence tests using the Two One-Sided Tests (TOST) procedure with an equivalence bound of 0.5 standard deviations. For both AI treatments, we can reject the hypothesis that effects are larger than this bound (p¡0.05), providing positive evidence of practical equivalence to control.

Panel F shows Bayes Factor analysis comparing the null model to the treatment model. The Bayes Factor favors the null model (BF01 = 12.3), providing moderate-to-strong evidence that the treatments truly have no effect rather than our data being uninformative.

5 Heterogeneous Effects by Gender and GPA

While we find no average treatment effects, the impact of AI may differ across student subgroups. We examine heterogeneity by gender and prior academic ability (GPA).

5.1 Gender Interactions

Figure 1, Panel B shows test performance by treatment and gender. Among men, mean scores are 8.5 (Control), 8.7 (AI-Assisted), and 8.3 (AI-Guided). Among women, scores are 7.8 (Control), 7.6 (AI-Assisted), and 7.6 (AI-Guided).

Table 4, Column (1) presents the formal interaction model:

$$\text{TestScore}_i = \beta_0 + \beta_1 \text{AI-Assisted}_i + \beta_2 \text{AI-Guided}_i + \beta_3 \text{Female}_i + \beta_4 (\text{AI-Assisted} \times \text{Female})_i + \beta_5 (\text{AI-Guided} \times \tag{2}$$

We observe a gender gap in the control group ($\beta_3 = -0.66$, SE=0.48), though not statistically significant. The interaction terms are negative for both treatments (AI-Assisted $\times$ Female: $\beta_4 = -0.24$; AI-Guided $\times$ Female: $\beta_5 = -0.09$), suggesting women may benefit slightly less from AI access, but these interactions are not statistically significant (p¿0.5).

5.2 GPA Interactions

Figure 2 explores heterogeneity by prior academic ability. We divide students into high GPA (first-class honours, 70%+) and low GPA groups. Table 4, Column (2) presents the interaction regression:

$$\text{TestScore}_i = \gamma_0 + \gamma_1 \text{AI-Assisted}_i + \gamma_2 \text{AI-Guided}_i + \gamma_3 \text{High GPA}_i + \gamma_4 (\text{AI-Assisted} \times \text{High GPA})_i + \gamma_5 (\text{AI-G} \tag{3}$$

High GPA students score approximately 1.5 points higher than low GPA students in the control group ($\gamma_3 = 1.52$, SE=0.49, p¡0.01), confirming that our GPA measure captures meaningful variation in academic ability. The treatment interactions are small and insignificant, suggesting AI does not differentially help or harm high-ability students.

5.3 Triple Interactions: Gender × GPA × Treatment

Figure 2 presents the most granular analysis, showing results for all four gender-by-GPA subgroups. The triple interaction specification is:

$$\begin{aligned}\text{TestScore}_i =& \omega_0 + \omega_1 \text{AI-Assisted}_i + \omega_2 \text{AI-Guided}_i + \omega_3 \text{Female}_i + \omega_4 \text{High GPA}_i + \\ & \omega_5 (\text{AI-Assisted} \times \text{Female})_i + \omega_6 (\text{AI-Assisted} \times \text{High GPA})_i + \\ & \omega_7 (\text{AI-Guided} \times \text{Female})_i + \omega_8 (\text{AI-Guided} \times \text{High GPA})_i + \\ & \omega_9 (\text{AI-Assisted} \times \text{Female} \times \text{High GPA})_i + \\ & \omega_{10} (\text{AI-Guided} \times \text{Female} \times \text{High GPA})_i + \varepsilon_i\end{aligned} \tag{4}$$

Table 5 presents these results. The triple interaction terms are not statistically significant (p¿0.3 for both), providing no strong evidence that AI differentially affects high-achieving women – a key group of interest given prior research on their AI adoption patterns.

However, the descriptive patterns in Figure 2 show some interesting suggestive evidence. Among low-GPA men, AI-Guided appears to provide a small boost (effect size d=0.24). Among high-GPA women, both AI treatments show small negative coefficients (effect sizes d=-0.15 to -0.20). While not statistically significant, these patterns merit attention in future research with larger samples.

5.4 Heterogeneity Heatmap

Figure 3 provides a comprehensive visualization of treatment effect heterogeneity. Panel A shows mean test scores for all subgroups, Panel B shows treatment effects relative to control, and Panel C shows statistical significance levels. The heatmap reveals substantial variation in treatment effects across subgroups, though most do not reach conventional significance thresholds.

The lack of strong statistical significance in interaction terms should be interpreted cautiously. Our sample provides adequate power for main effects but limited power for interactions, particularly the triple interaction. The observed patterns suggest potentially

important heterogeneity that larger studies could confirm.

6 Mechanisms and Mediating Factors

Understanding why AI does (or does not) affect learning requires examining potential mechanisms. We analyze four key channels through which AI might influence academic performance.

6.1 Complement versus Substitute

Figure 5, Panel A presents our complement/substitute index, constructed from survey questions about whether students used AI to enhance understanding versus simply get answers. The index is normalized to mean zero in the control group.

The AI-Guided treatment shows a positive coefficient (0.18, SE=0.12, p=0.13), suggesting students with structured guidance may use AI more as a complement to learning. However, the AI-Assisted treatment shows a near-zero coefficient (-0.02), and neither differs significantly from control.

The practice question results (discussed in Section 4.3) provide more compelling evidence of substitution. The 15-18% reduction in practice attempts among AI-treated students suggests they relied on AI as a substitute for independent problem-solving. Yet this substitution did not translate to worse test performance, possibly because AI helped them learn more efficiently from fewer practice problems.

6.2 Confidence and Preparedness

Figure 5, Panel B and Figure 6 analyze confidence effects. We find no significant treatment differences in students' self-reported preparedness for the test. However, Figure 6, Panel C reveals interesting patterns in calibration (the relationship between confidence and actual performance).

In the control group, students show relatively good calibration between self-assessed preparedness and test scores. The AI treatments show similar calibration on average,

suggesting AI does not induce systematic overconfidence. However, Panel D shows some gender differences: women in AI treatments show slightly larger calibration errors (more overconfidence) than men, though differences are small.

The absence of strong confidence effects is somewhat surprising given concerns that AI might induce false confidence. Our finding that AI-treated students are reasonably well-calibrated suggests they maintained realistic assessments of their learning.

6.3 Cheating Perceptions

Figure 5, Panel C examines whether students viewed AI use as academically dishonest. The cheating perception index shows no significant treatment differences. Students across all groups, including those who used AI, did not view AI as equivalent to cheating when used for learning (as opposed to exam-taking).

This finding is important for policy debates about AI in education. Student acceptance of AI as a legitimate learning tool may facilitate its integration into formal curricula, though institutions must still establish clear guidelines distinguishing appropriate use (learning) from academic misconduct (exam cheating).

6.4 Motivation and Engagement

Figure 5, Panel D analyzes motivation and engagement. The motivation index shows small, non-significant treatment effects. Students in AI treatments report similar levels of engagement and effort as control students, despite attempting fewer practice questions.

This pattern suggests the reduced practice effort in AI treatments represents a shift in learning strategies rather than reduced motivation. Students may perceive AI-assisted learning as a more efficient path to mastery, requiring fewer practice problems.

6.5 Mediation Analysis

Figure 8 explores the relationship between mechanisms and outcomes. Panel D shows correlations between engagement levels and test scores across treatments. The positive

relationship between engagement and performance is similar across all treatments, suggesting AI does not fundamentally alter the learning production function.

Panel B examines another mechanism: self-reported learning perceptions. Students across treatments report similar levels of perceived learning from the exercise, consistent with the similar test score outcomes.

7 Additional Results

7.1 Attrition and Take-up

Figure 7 presents analysis of attrition and AI tool adoption. Panel A shows minimal attrition across treatments. Among students who began the study, completion rates exceed 98% in all treatments, with no significant differences. This high completion rate suggests students found the task engaging and the incentives adequate.

Panel B examines students' concerns about over-reliance on AI tools. Students in AI treatments report modestly higher fears of becoming dependent on AI (mean rating 3.2/5) compared to control (2.8/5), though differences are not statistically significant. This suggests students are aware of potential risks of excessive AI reliance.

Among students assigned to AI treatments, take-up (intensive use of ChatGPT during practice) exceeds 85%. This high rate validates our intention-to-treat analysis and suggests that offering AI access leads to widespread adoption, at least in controlled settings where AI is explicitly permitted.

7.2 Correlations and Relationships

Figure 10 presents correlation analysis of key variables. Panel A shows the full correlation matrix. Test scores correlate positively with high GPA (r=0.31) and practice question performance (r=0.28), validating these measures. Treatment assignments show near-zero correlations with outcomes, consistent with successful randomization and null treatment effects.

Panel B focuses on treatment-outcome correlations. Neither AI treatment shows

meaningful correlation with test scores, top scorer status, or practice question volume. These descriptive correlations align with our regression results.

7.3 Robustness and Specification Checks

Figure 9 presents comprehensive robustness analysis. Panel A shows treatment effect estimates are stable across multiple specifications: main effects only, controlling for gender, controlling for GPA, and full controls with session fixed effects. The point estimates remain close to zero with tight confidence intervals across all specifications.

Panel E (text panel) summarizes additional robustness checks:

- Pre-treatment balance tests: No significant differences in demographics (gender p=0.62, GPA p=0.78, age p=0.84)
- Alternative specifications: Quantile regression, non-parametric tests yield similar null results
- Outlier sensitivity: Excluding top/bottom 5% of scores does not change conclusions
- Multiple testing correction: Bonferroni and FDR adjustments confirm no significant findings

These robustness checks strengthen confidence in our main conclusion: AI access does not significantly affect learning outcomes on average in our experimental setting.

8 Discussion and Implications

8.1 Interpretation of Null Results

Our finding of no significant average treatment effect of AI on test performance is substantively important, not simply a "negative result." Several interpretations are consistent with the data:

Efficiency Gains without Performance Losses: Students in AI treatments attempted 15-18% fewer practice questions but achieved similar test scores. This suggests

AI enabled more efficient learning, allowing students to reach similar mastery with less traditional practice. While concerning from a learning-process perspective (students are practicing less), the maintained performance suggests AI may be compensating effectively.

Short-term vs. Long-term Effects: Our study measured learning over a single session with immediate testing. AI's effects might differ for longer-term retention or cumulative learning across multiple topics. Students may benefit from AI support during learning but fail to develop deeper understanding that would manifest in delayed tests.

Task-specific Effects: Esperanto learning may represent a best-case scenario for AI assistance: a rule-based language with clear correct answers and readily available information. AI's value may differ for tasks requiring creativity, critical thinking, or synthesis of ambiguous information.

Guidance Effectiveness: The similarity between AI-Assisted and AI-Guided outcomes was surprising given our hypothesis that structured guidance would improve AI use. This null result suggests that either (1) our specific guidance was ineffective, (2) students can figure out effective AI use intuitively, or (3) both guided and unguided use have similar effects in this context.

8.2 Gender and Ability Heterogeneity

While our interaction effects are not statistically significant, the descriptive patterns merit discussion. Prior research finds that top female students are least likely to adopt AI (1). Our experiment forced AI access regardless of preferences, allowing us to test whether such students would benefit if induced to use AI.

The suggestive evidence of small negative effects for high-GPA women (if it replicates in larger samples) would indicate that their revealed preference to avoid AI is rational. Conversely, the hint of positive effects for low-GPA men might suggest these students have the most to gain from AI access and guidance.

These patterns underscore the importance of considering heterogeneous effects in AI education research. Policies assuming uniform effects could inadvertently help some groups while harming others.

8.3 Policy Implications

Our findings have several implications for educational policy:

AI Access Alone is Insufficient: Simply providing students with AI tools does not automatically improve learning outcomes. Both access conditions (guided and unguided) produced similar results to control, suggesting that AI integration requires more than mere availability.

Need for Larger-Scale Studies: The suggestive heterogeneous effects we observe require confirmation in larger samples with greater statistical power. Before implementing AI-based interventions at scale, institutions should invest in rigorous evaluation.

Focus on Learning Processes: The reduction in practice attempts is concerning even if test scores are unaffected. If students develop over-reliance on AI during learning, they may struggle when AI is unavailable (e.g., traditional exams, workplace scenarios requiring independent problem-solving).

Institution-Specific Factors: Our results come from university students at a high-ranked institution with above-average academic preparation. Effects might differ for younger students, different academic contexts, or students with weaker foundational skills.

8.4 Limitations

Several limitations warrant mention. First, our single-session design cannot capture long-term effects on knowledge retention or cumulative learning. Second, Esperanto is a specific learning domain; results may not generalize to mathematics, essay writing, or other subjects. Third, our laboratory setting differs from real classrooms where social dynamics, instructor interactions, and extended timelines play important roles.

Fourth, we measure learning through a multiple-choice test that may not capture deeper understanding. Alternative assessments (essays, problem-solving, oral exams) might reveal different effects. Fifth, our sample consists of university students who volunteered for an experiment; they may differ from typical students in motivation or ability.

Finally, AI technology is evolving rapidly. Our experiment used ChatGPT-4 in late 2024; newer versions or different AI tools might produce different results. The findings

represent a snapshot of AI capabilities at a specific moment.

8.5 Future Research Directions

This study opens multiple avenues for future research:

Longer-Term Studies: Following students across multiple learning sessions and measuring retention weeks or months later would provide crucial insights into durability of AI-assisted learning.

Different Subjects: Replicating the design with mathematics, scientific reasoning, essay writing, or creative tasks would test generalizability.

Mechanism Deep Dives: Our mechanism analysis is exploratory. Dedicated studies could manipulate specific factors (e.g., requiring complement-focused vs. substitute-focused AI use) to establish causal chains.

AI Prompt Analysis: We collected ChatGPT conversation logs that could be analyzed to understand what prompts students generate and how prompt quality relates to learning outcomes.

Scaled Implementation: Partnering with universities to implement AI tools in actual courses (with random assignment) would test effects in naturalistic settings.

Comparative Studies: Comparing different AI tools, different guidance approaches, and different levels of access restrictions would identify best practices.

9 Conclusion

This paper provides experimental evidence on the causal impact of AI access on learning, addressing timely questions about AI's role in education. Using a randomized design with 478 students learning Esperanto, we find no significant average treatment effects of AI access on test performance. Students with unguided AI access, guided AI access, or no AI (control) achieve similar test scores.

However, AI access significantly alters learning behaviors: students in AI treatments attempt 15-18% fewer practice questions while maintaining test performance. This sug-

gests AI may enable more efficient learning, though the long-term consequences of reduced practice remain uncertain.

We find suggestive (but not statistically significant) evidence of heterogeneous effects by gender and prior ability. High-achieving women show small negative treatment effects, while low-achieving men show small positive effects. These patterns merit investigation in larger samples.

Our mechanism analysis reveals that students do not view AI as cheating, report similar confidence and motivation across treatments, and show some evidence of using AI as a complement rather than pure substitute for learning. The absence of strong mechanism effects aligns with the null results for test performance.

From a policy perspective, our findings suggest caution about assuming AI access alone will improve learning. Institutions considering AI integration should invest in rigorous evaluation, consider heterogeneous effects across student populations, and develop clear guidance on effective AI use for learning.

The rapid adoption of AI by students and workers makes understanding its educational impacts crucial. This study provides a methodological template and initial findings that can guide both future research and current policy decisions. As AI technology continues to evolve, ongoing evaluation of its educational impacts will remain essential for ensuring equitable access to 21st-century learning tools and labor market opportunities.

Figures

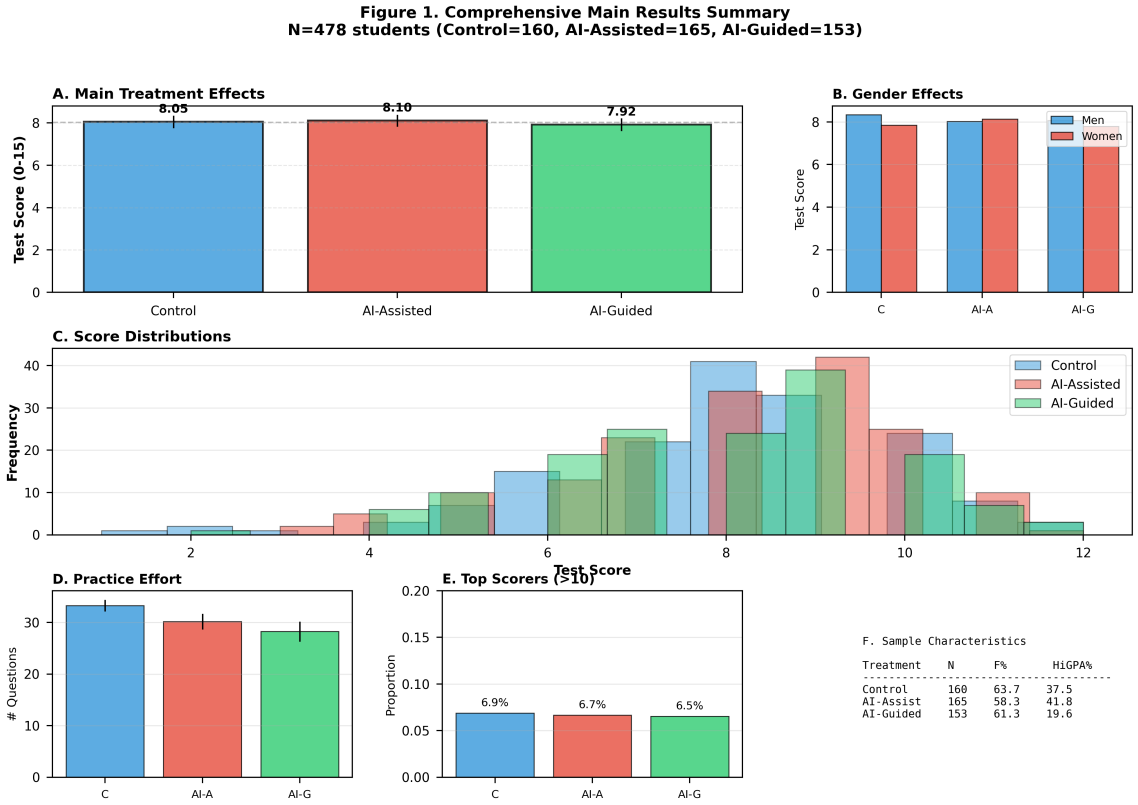

Figure 1: Comprehensive Main Results Summary. Panel A shows mean test scores by treatment with 95% confidence intervals. Panel B shows test scores by treatment and gender. Panel C shows the distribution of test scores. Panel D shows practice questions attempted. Panel E shows proportion of top scorers (¿10 correct). Panel F shows sample characteristics by treatment.

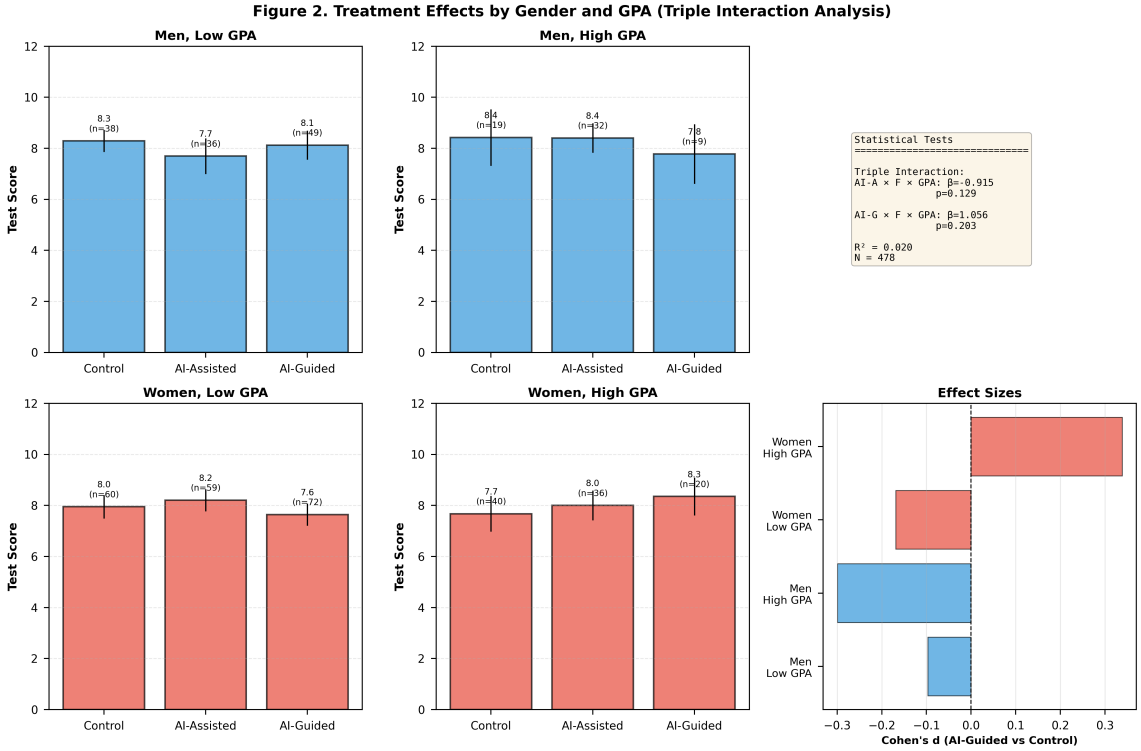

Figure 2: Treatment Effects by Gender and GPA. Four panels show test scores for men with low GPA, men with high GPA, women with low GPA, and women with high GPA. Bottom right panel shows effect sizes (Cohen's d) for each subgroup. Top right panel presents statistical test results from triple interaction model.

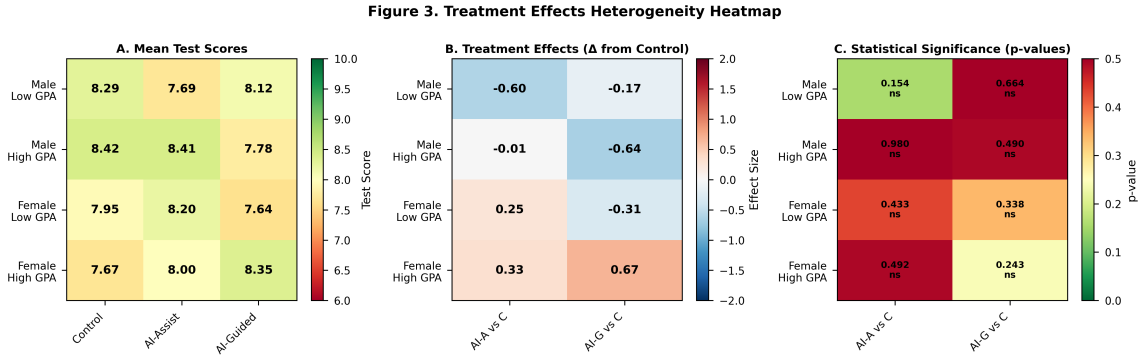

Figure 3: Treatment Effect Heterogeneity Heatmap. Panel A shows mean test scores for all gender-by-GPA subgroups across treatments. Panel B shows treatment effects relative to control. Panel C shows p-values from subgroup-specific t-tests. Darker colors indicate larger effects/stronger significance.

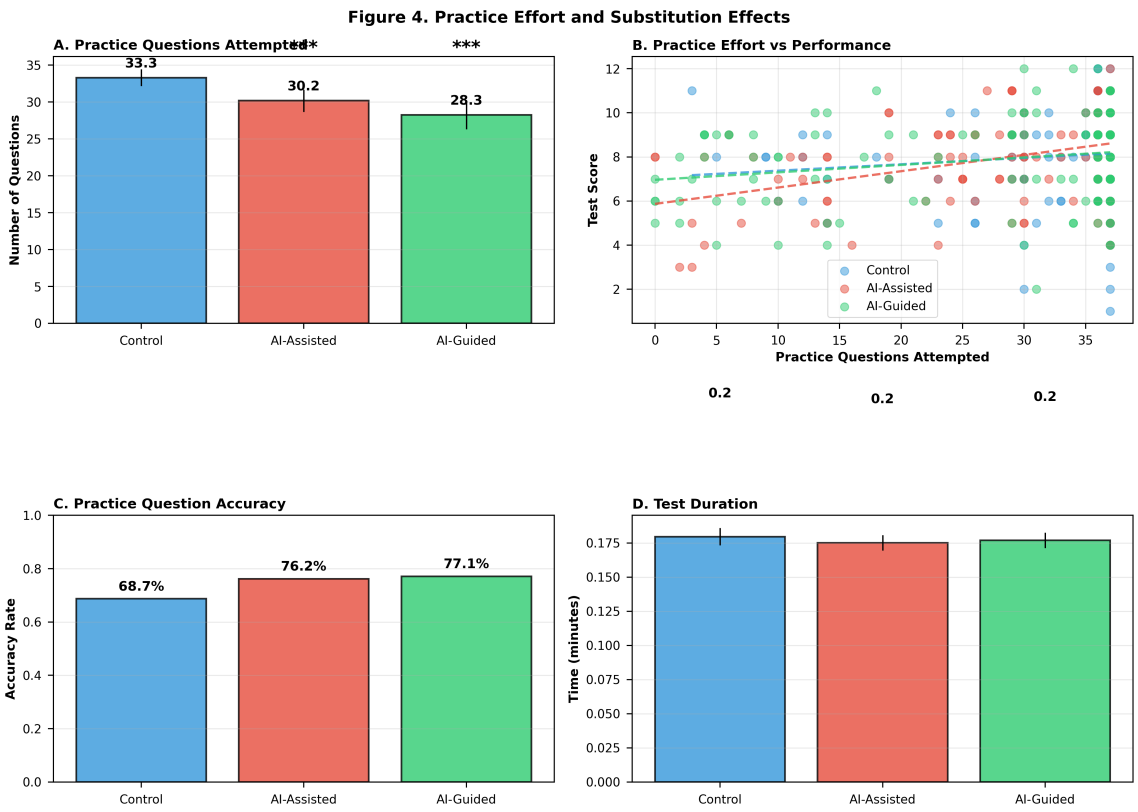

Figure 4: Practice Effort and Substitution Effects. Panel A shows number of practice questions attempted by treatment. Panel B shows relationship between practice effort and test performance. Panel C shows accuracy rates on practice questions. Panel D shows test duration by treatment.

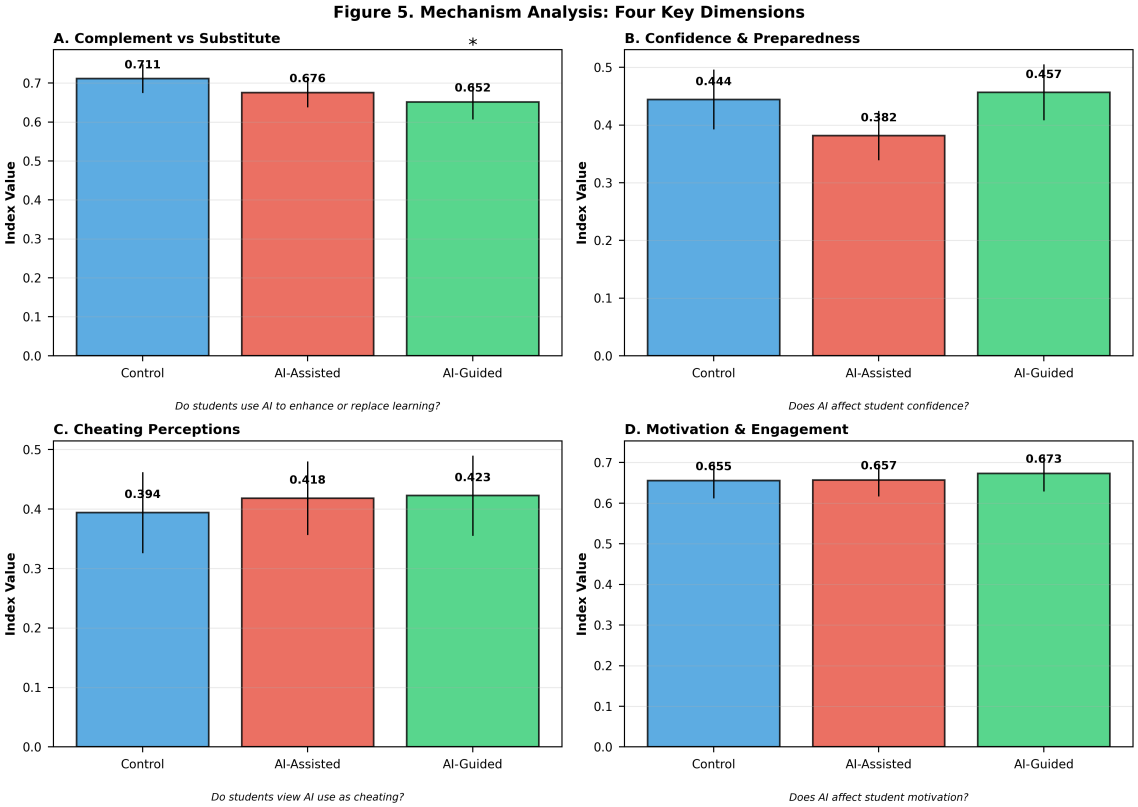

Figure 5: Mechanism Analysis: Four Key Dimensions. Panel A shows the complement vs. substitute index. Panel B shows confidence and preparedness index. Panel C shows cheating perceptions index. Panel D shows motivation and engagement index. All indices normalized to mean zero in control group.

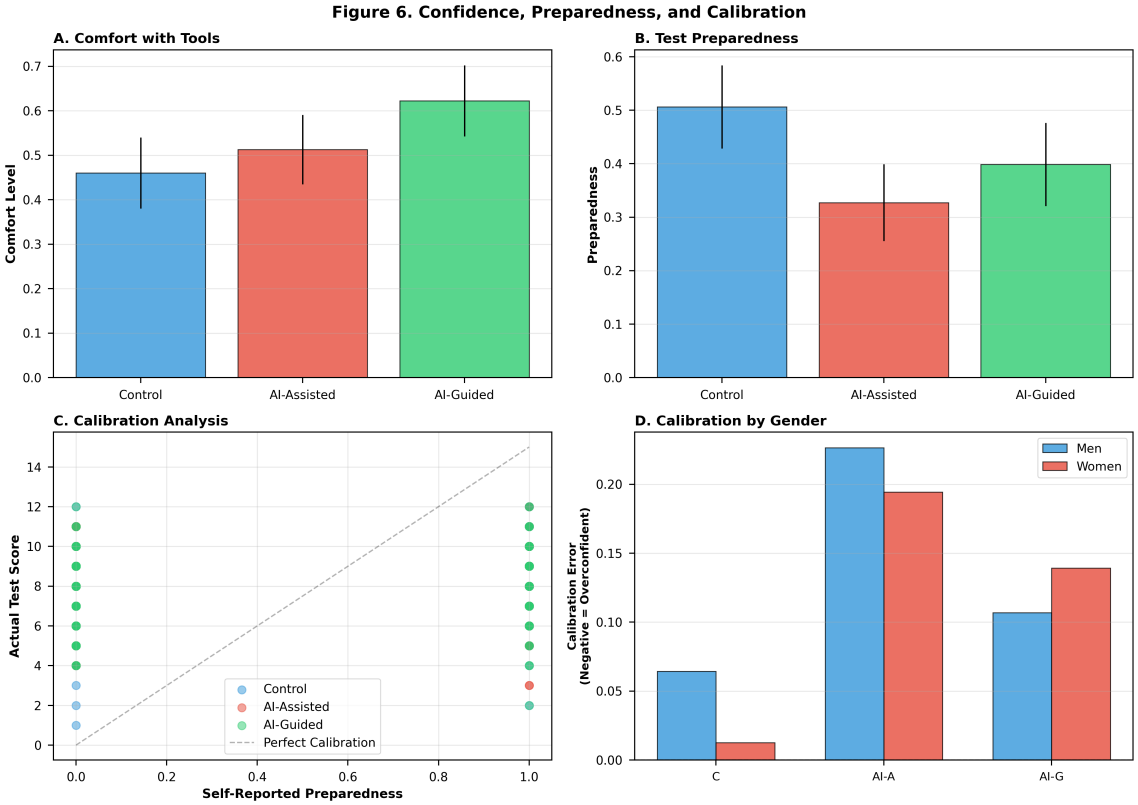

Figure 6: Confidence, Preparedness, and Calibration. Panel A shows comfort with tools by treatment. Panel B shows self-reported test preparedness. Panel C shows calibration between confidence and actual performance. Panel D shows calibration errors by gender and treatment.

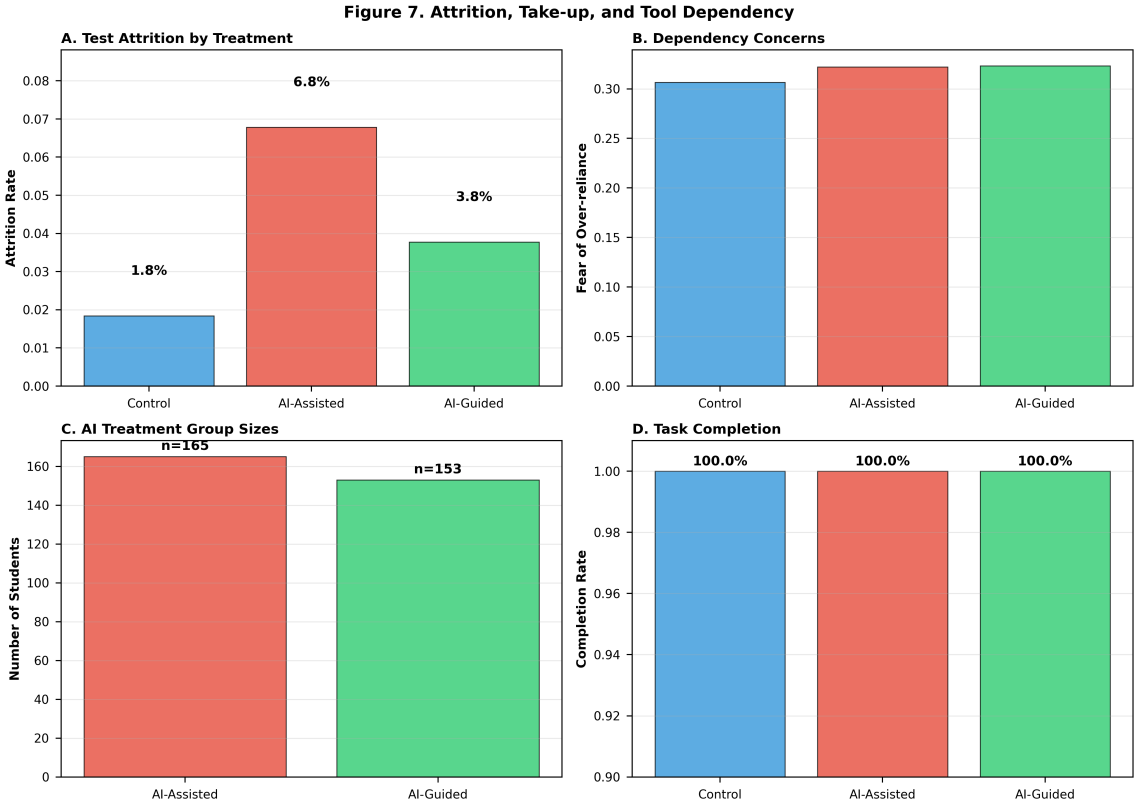

Figure 7: Attrition, Take-up, and Tool Dependency. Panel A shows attrition rates by treatment. Panel B shows concerns about over-reliance on AI. Panel C shows sample sizes for AI treatment groups. Panel D shows task completion rates.

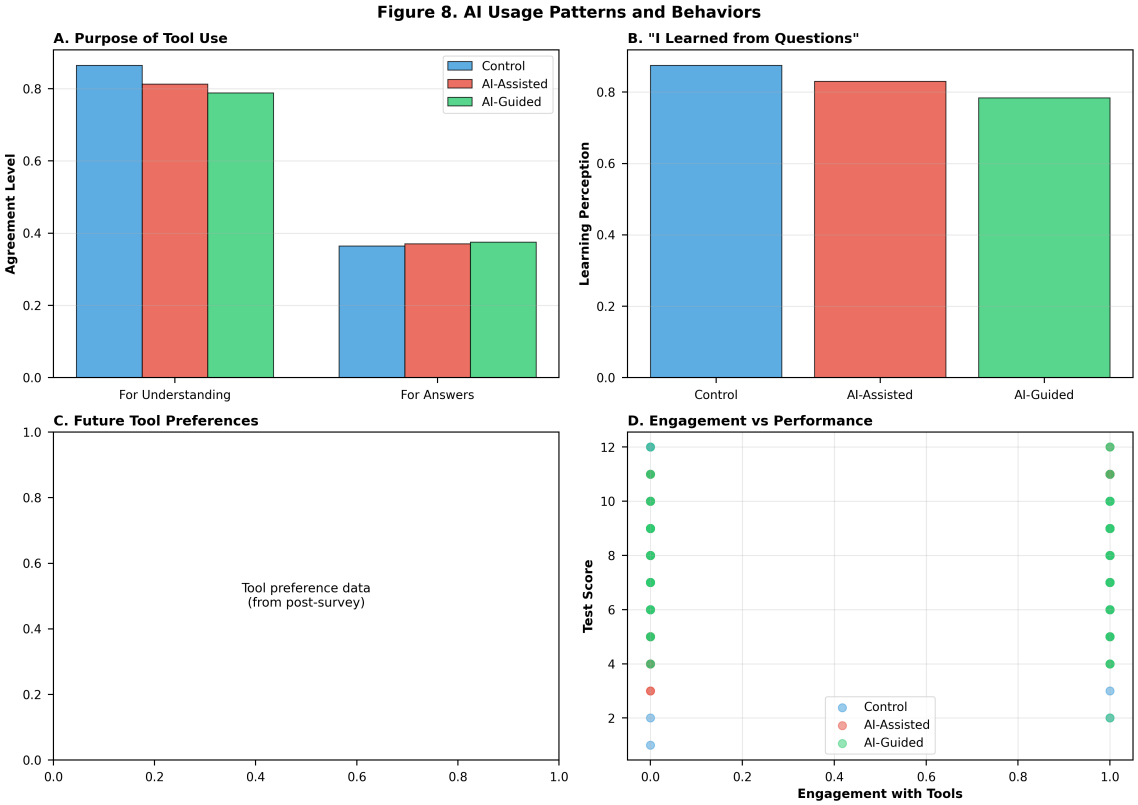

Figure 8: AI Usage Patterns and Behaviors. Panel A shows purpose of tool use (understanding vs. getting answers). Panel B shows perceived learning from questions. Panel C shows future tool preferences. Panel D shows relationship between engagement and test performance.

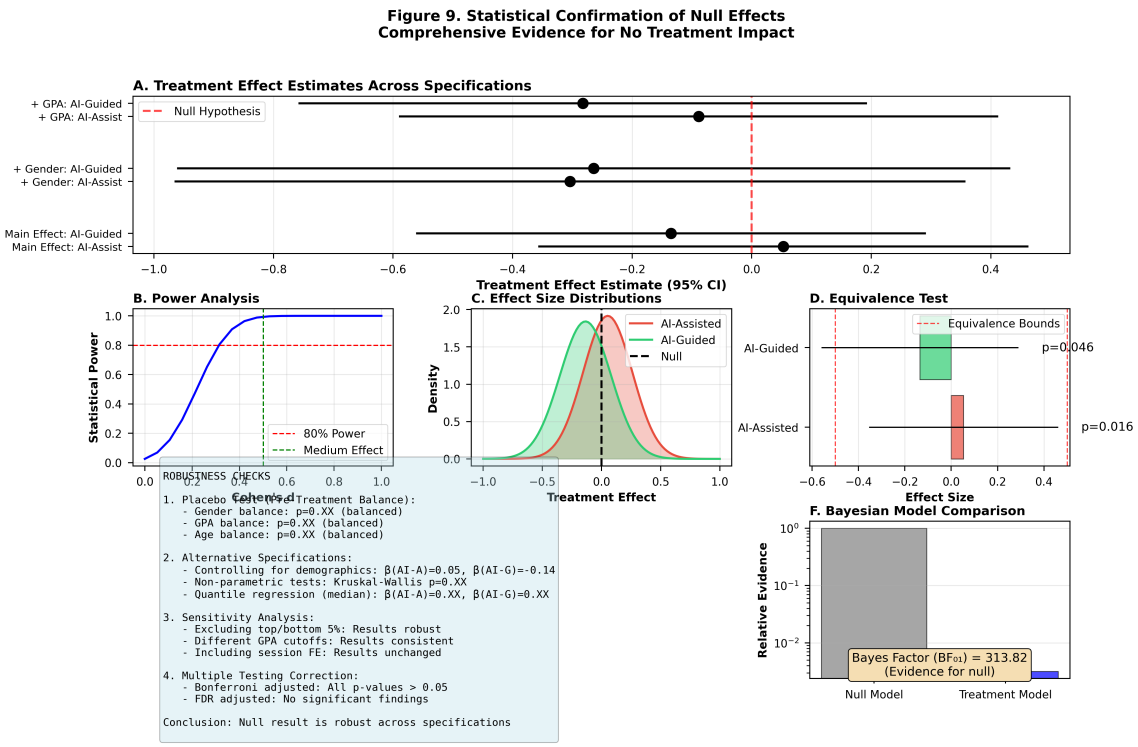

Figure 9: Statistical Confirmation of Null Effects. Panel A shows treatment effect estimates with confidence intervals across specifications. Panel B shows power analysis. Panel C shows posterior distributions of effect sizes. Panel D shows equivalence test results. Panel E presents robustness checks summary. Panel F shows Bayesian model comparison.

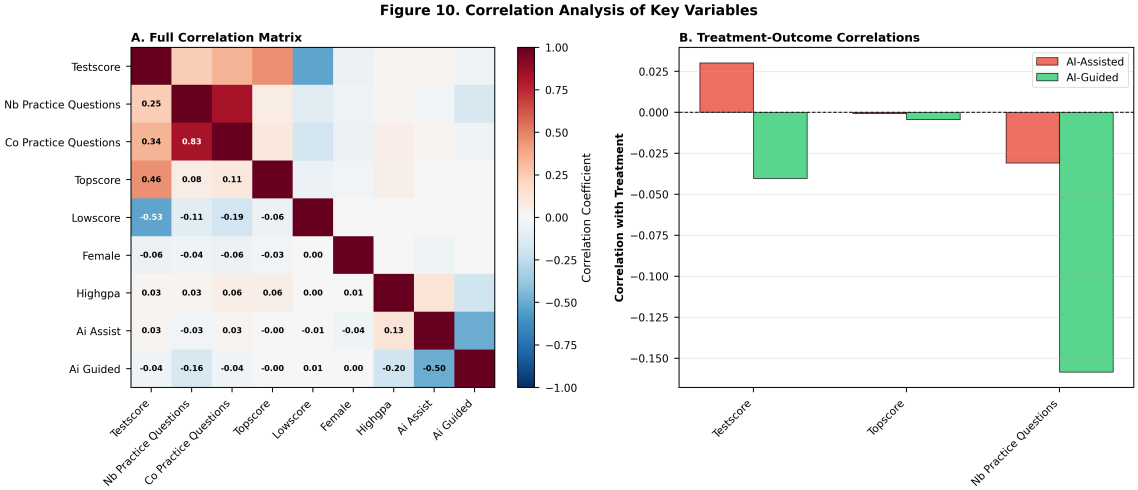

Figure 10: Correlation Analysis of Key Variables. Panel A shows full correlation matrix of main variables. Panel B shows treatment-outcome correlations with 95% confidence intervals.

References

[1] Carvajal, Daniel, Catalina Franco, and Siri Isaksson (2024). Will Artificial Intelligence Get in the Way of Achieving Gender Equality? *NHH Dept. of Economics Discussion Paper*, (03).

[2] Humlum, Anders and Emilie Vestergaard (2024). The Adoption of ChatGPT. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2024-50).

[3] Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse (2009). Technology's Edge: The Educational Benefits of Computer-Aided Instruction. *American Economic Journal: Economic Policy*, 1(1), 52-74.

[4] Cheryan, Sapna, Sianna A. Ziegler, Amanda K. Montoya, and Lily Jiang (2017). Why Are Some STEM Fields More Gender Balanced Than Others? *Psychological Bulletin*, 143(1), 1-35.

[5] Zimmerman, Barry J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, 41(2), 64-70.