Students' Achievement in Mathematics Through AI-Assisted Teaching: A Quasi-Experimental Study

[1] Ashish Sharma, [2] Dr. Satyendra Gupta
[1]Ph.D. (Education) Research Scholar, Galgotias University, Greater Noida, UP, India
Email- ashashish.sharma16@gmail.com
[2]Professor & Dean, School of Education Galgotias University, Greater Noida, UP, India
Email- satyendra.edn@gmail.com

HOW TO CITE:
Ashish Sharma, Satyendra Gupta (2026). Students' Achievement In Mathematics Through Ai-Assisted Teaching: A Quasi-Experimental Study. International Journal of Special Education, 41(18s), 1479-1486

ABSTRACT:
The use of Artificial Intelligence (AI) in education is expanding, providing opportunities for individual, adaptive, and learner-centred mathematics teaching and learning. The present study was conducted in a school of Rewari in Haryana, India to know the effect of using AI in teaching on students' achievement in mathematics. The quasi-experimental pretest – posttest control group design was used. The subjects of the study were 83 students who were distributed into two groups; experimental group (42) and control group (41) from one School. The experimental group was instructed in mathematics using AI for 8 weeks and the control group was instructed in mathematics using traditional methods. The Achievement Test in Mathematics (ATM) was a 40-item researcher-developed mathematics achievement test used to assess students' mathematics achievement. Descriptive and inferential statistical techniques were employed to analyse the data. The results showed that students in the AI-assisted mathematics group had a better mathematics achievement level at the post-test stage than the students in the traditional mathematics group. The results, however, must be interpreted with caution because differences in mathematics achievement were found to exist between the groups prior to the intervention; thus, the study cannot conclude that AI-assisted teaching has a causal effect on mathematics achievement. However, the results indicate that there is potential for using AI-assisted instruction as one of the effective measures to aid mathematics learning and improve students' achievement.

Keywords: Artificial Intelligence, AI-Assisted Teaching, Mathematics Achievement, School Education, Quasi-Experimental Study, ANCOVA, Achievement Test

1. Introduction

The attainment in maths is one of the key measures of school quality and one of the most elusive to raise. In a traditional whole-class setting, the teacher may find it difficult to balance students' diverse prior knowledge, speed of work, and misconceptions when teaching mathematics, especially in low-resourced classrooms. One potential solution to this variability is the use of Artificial Intelligence (AI) to sequence content adaptively, give feedback in real time, and have students practice what they have learned individually, or in small groups of students, at a level a single teacher might not be able to do

(Zawacki-Richter et al., 2019). AI in mathematics education has been systematically reviewed and can be found to have a growing evidence base that is methodologically inconsistent, with intelligent tutoring systems, adaptive practice platforms, and AI generated feedback tools being reported at all levels of mathematics instruction, primary, secondary and tertiary (Mohamed et al., 2022). Studies that have conducted meta-analyses on computer- and AI-based tutoring indicate that such systems can boost achievement above that of no supplemental instruction, although the magnitude of the instructional effect has been found to vary widely depending on the level of accomplishment provided, the amount of time spent using the computer system, and the comparison condition (Ma et al., 2014; VanLehn, 2011). However, recent studies have found more mixed results in the context of resource-limited, rural schools, and mathematics achievement – with positive effects found in some contexts, but not others (Khazanchi, Di Mitri, & Drachsler, 2025).

This means that the impact of AI-based systems on mathematics achievement is not necessarily uniform across contexts and requires local evaluation. AI-backed platforms are being touted as a viable alternative to classroom instruction in schools that frequently have high student-to-teacher ratios and large classes. Evidence of their impact on measured achievement, however, is limited and is done at a local level with high methodological risk because of the tendency for school-based quasi-experimental designs to suffer from methodological problems. The present study attempts to address this gap by implementing an AI-assisted teaching–learning process to enhance mathematics achievement among Class II students in a private school in Rewari, Haryana. A pre-test–post-test control group quasi-experimental design was employed, with appropriate statistical procedures used to account for the pre-existing difference in pre-test scores between the experimental and control groups.

2. Review of Related Literature

Zawacki-Richter et al. (2019), in a systematic review of AI applications in higher education, found that most published work emphasizes system design and performance rather than pedagogical grounding, and called for more classroom-based evaluations that connect AI tools to learning outcomes and involve educators directly in design and evaluation. Mohamed et al. (2022), reviewing AI applications specifically in mathematics education, similarly reported that intelligent tutoring systems, automated assessment, and adaptive learning platforms dominate the literature, with achievement gains reported inconsistently across studies of varying methodological rigor.

Two influential meta-analyses provide effect-size benchmarks against which classroom studies can be compared. VanLehn (2011) compared human tutoring, intelligent tutoring systems, and no-tutoring conditions across 28 evaluation studies and reported effect sizes of approximately $d = 0.75$ for step-based intelligent tutoring relative to no tutoring, noticeably lower than the 2.0 benchmark historically attributed to one-to-one human tutoring. Ma et al. (2014), in a meta-analysis of intelligent tutoring systems, reported a mean effect of $g \approx 0.66$ relative to conventional instruction, with larger effects for systems used over longer periods and for well-implemented, curriculum-aligned tools. These benchmarks suggest that AI-based tutoring effects, while frequently positive, are typically small-to-moderate and sensitive to dosage and implementation quality — a pattern consistent with the modest, non-significant adjusted effect observed in the present study.

Closer to the present context, Khazanchi et al. (2025) examined the effect of an AI-based adaptive system (Edmentum Exact Path) on mathematics achievement among rural K-12 students using a quasi-experimental design, reporting improvements in achievement alongside gains in student engagement, while also noting that a quasi-experimental design without random assignment limits causal interpretation

— a limitation acknowledged in the present study as well.

A further body of methodological literature is directly relevant to the analytic approach adopted here. Dimitrov and Rumrill (2003) discuss the specific problem of comparing intact, non-randomly assigned groups on pre-test and post-test measures, noting that raw post-test comparisons can be misleading when groups differ at baseline, and recommending analysis of covariance (ANCOVA) with the pre-test as covariate — or, equivalently, analysis of gain scores — as more defensible alternatives. This methodological guidance directly informed the analytic strategy of the present study, in which both a raw post-test comparison and a pre-test-adjusted (ANCOVA) comparison are reported and contrasted.

Taken together, the literature suggests that (a) AI-assisted instruction is associated with positive, though variable, effects on mathematics achievement; (b) reported effect sizes are typically small-to-moderate rather than large; and (c) methodologically rigorous evaluation of such interventions in non-randomized school settings requires explicit statistical control for baseline non-equivalence. The present study was designed with these considerations in mind.

3. Statement of the Problem

The study is titled "Effect of Artificial Intelligence-Assisted Teaching on Students' Achievement in Mathematics." It investigates whether an AI-assisted instructional intervention, delivered over more than eight weeks to primary school students in Rewari, Haryana, produces a measurable improvement in mathematics achievement relative to regular classroom instruction, and whether any observed difference persists after statistically accounting for pre-existing differences between the intervention and comparison groups.

4. Objectives of the Study

O_{01}. To compare the pre-test mathematics achievement of the experimental and control groups prior to the intervention.

O_{02}. To compare the post-test mathematics achievement of the experimental group (AI-assisted teaching) and the control group (regular instruction).

O_{03}. To compare the pre-test to post-test gain in mathematics achievement within each group.

O_{04}. To examine the post-test difference between groups after statistically controlling pre-test achievement using ANCOVA.

5. Hypotheses

H_{01}: There is no significant difference between the pre-test mathematics achievement scores of the experimental and control groups.

H_{02}: There is no significant difference between the post-test mathematics achievement scores of the experimental and control groups.

H_{03}: There is no significant difference between the pre-test and post-test mathematics achievement scores within each group.

H_{04}: There is no significant difference between the adjusted post-test mathematics achievement scores of the experimental and control groups after controlling pre-test scores.

6. Method

6.1 Research Design

This study was of a quasi-experimental type with a non-equivalent group, pre-test–post-test control group design. No attempt was made to randomize individual students into groups; instead, intact classroom groups were used, which is often the case in school-based intervention research. Due to this design, pre-test adjusted (ANCOVA) comparisons as well as between-group comparisons of raw data are reported because this design does not provide baseline equivalence between groups.

6.2 Sample

The sample consisted of 83 Class II students from one private school in Rewari, Haryana. Of these, 42 students (23 boys and 19 girls) constituted the experimental group, while 41 students (23 boys and 18 girls) constituted the control group. All available Class II students in the selected school were included in the study.

The students were assigned to experimental and control groups based on research convenience rather than random assignments. The gender distribution was broadly comparable across the two groups.

6.3 Tool Used

Mathematics achievement was assessed using a researcher-developed Achievement Test in Mathematics (ATM). Initially, 97 multiple-choice questions (MCQs) were prepared covering four content areas relevant to Class II mathematics: Shapes, Numbers, Mathematical Operations, and Patterns. The preliminary test was reviewed by subject-matter and language experts to assess the relevance, clarity, language, and age-appropriateness of the items. The 97-item preliminary test was subsequently administered to 448 Class II students for item analysis. The difficulty index and discrimination index were calculated to evaluate the quality and suitability of the items. Based on the expert review and item analysis, 40 items were finalized for the Achievement Test in Mathematics (ATM), which was used as the final instrument to assess students' mathematics achievement in the present study.

6.4 Intervention Procedure

Both groups administered the Achievement Test in Mathematics as a pre-test prior to the intervention. The experimental group then received AI-assisted mathematics instruction for a period of more than eight weeks, while the control group continued to receive regular (conventional) classroom instruction covering the same curricular content over the same period. Following the intervention period, both groups were administered for the same test as a post-test.

a. Experimental Group: AI-Assisted Teaching

The students in the experimental group were taught mathematics using the Chimple Learning and Chimple Class applications supported by AI learning assistant. The intervention took place for 10–15 minutes per day, with students accessing a parent's cellphone and being supervised by a parent while completing the lessons. Alongside the AI-based sessions, additional self-learning tasks and maths homework were also given to be completed at home, to enhance practice and continued engagement. The AI-based learning environment provided adaptive learning activities, immediate feedback, interactive problem-solving opportunities, and gamified reinforcement. During the intervention, the teacher gave guidance and support when needed, and the AI-assisted applications were the main medium for the students' individual practice and learning.

b. Control Group: Conventional Classroom Teaching

The Control Group's pupils were taught mathematics using traditional teaching techniques. The instructor employed chalkboard-based explanation, textbook exercises, oral problem-solving, and demonstrations. Throughout the intervention, both groups received equal instructional time, covered the same mathematical topics, and adhered to the same syllabus.

6.5 Statistical Treatment

Descriptive statistics (means, standard deviations) were computed for pre-test and post-test scores in each group. Shapiro Wilk tests and Levene's test were used to check the normality and homogeneity-of-variance assumptions underlying parametric comparisons. Independent-samples t-tests (with Welch's correction applied where Levene's test indicated unequal variances) were used to compare groups on pre-test scores, post-test scores, and gain scores. Paired-samples t-tests were used to compare pre-test and post-test scores within each group. Analysis of covariance (ANCOVA), with the pre-test score entered as covariate and post-test score as the dependent variable, was used to compare adjusted post-test means between groups; the homogeneity-of-regression-slopes assumption was tested prior to interpreting the ANCOVA. Mann–Whitney U tests were additionally computed as a non-parametric robustness check given the non-normality detected in one group's pre-test distribution. Cohen's d, Cohen's dz (for paired designs), and partial eta-squared (η^2p) were computed as effect-size indices. The significance level was set at α = .05 throughout.

7. Results

Table 1. Descriptive Statistics for Pre-test and Post-test Mathematics Achievement Scores

Group	N	Pre-test M (SD)	Post-test M (SD)	Mean Gain (SD)
Experimental	42	13.21 (4.61)	23.12 (4.50)	9.90 (3.15)
Control	41	10.88 (5.31)	20.44 (6.73)	9.56 (6.66)

Table 2. Assumption Checks: Normality (Shapiro–Wilk) and Homogeneity of Variance (Levene's Test)

Measure	Exp. W (p)	Ctrl. W (p)	Levene's F (p)
Pre-test	0.914 (0.004)	0.975 (0.504)	1.55 (0.216)
Post-test	0.952 (0.075)	0.952 (0.082)	10.81 (0.001)

The pre-test distribution of the experimental group departed significantly from normality (p = 0.004), and Levene's test indicated significant heterogeneity of variance for the post-test (p = 0.001). These violations do not invalidate the t-test and ANCOVA results reported below. Welch's correction was applied where variances were unequal, and Mann–Whitney U tests are reported alongside parametric tests as a robustness check but they do warrant caution in interpretation and are discussed further in the Limitations section.

Table 3. Between-Group Comparison of Pre-test (Baseline) Scores

Test	Statistic	df	p	Effect size
Independent t-test (Student)	t = 2.14	81	0.035	d = 0.47
Independent t-test (Welch)	t = 2.14	78.88	0.036	d = 0.47
Mann–Whitney U	U = 1101.5	—	0.028	—

The experimental group scored significantly higher than the control group at pre-test, confirming that the two groups were not equivalent in mathematics achievement before the intervention began (H01 rejected). This baseline inequality is the central methodological complication that the ANCOVA reported in Table 6 is designed to address.

Table 4. Between-Group Comparison of Raw Post-test Scores

Test	Statistic	df	p	Effect size
Independent t-test (Student)	t = 2.14	81	0.036	d = 0.47
Independent t-test (Welch)	t = 2.13	69.58	0.037	d = 0.47
Mann–Whitney U	U = 1032.5	—	0.119	—

The raw post-test comparison favoured the experimental group and reached statistical significance under both the Student's t and Welch-corrected t (H02 rejected on raw scores), with a moderate effect size (d = 0.47) identical in magnitude to the pre-test difference. Notably, the non-parametric Mann–Whitney U test — which does not assume normality — did not reach significance for the post-test comparison (p = .119), foreshadowing the ANCOVA result below.

Table 5. Within-Group Pre-test to Post-test Comparison (Paired-samples t-test)

Group	Mean Gain (SD)	t	df	p	dz
Experimental	9.90 (3.15)	20.41	41	< .001	3.15
Control	9.56 (6.66)	9.20	40	< .001	1.44

Both groups showed statistically significant and substantial improvement from pre-test to post-test (H03 rejected for both groups), confirming that learning occurred under both instructional conditions. The independent-samples comparison of gain scores was not significant, t(81) = 0.30, p = .763, d = 0.07 — that is, the amount of improvement did not differ meaningfully between the AI-assisted and conventionally taught groups, even though the experimental group's within-group effect size (dz = 3.15) numerically exceeded that of the control group (dz = 1.44), a difference attributable in part to the larger score variability in the control group.

Table 6. Analysis of Covariance (ANCOVA) of Post-test Scores by Group, Controlling for Pre-test Scores

Source	SS	df	MS	F	p	η^2p
Pre-test (covariate)	879.82	1	879.82	37.37	< 0.001	0.318
Group	30.17	1	30.17	1.28	0.261	0.016
Error	1883.53	80	23.54			
Total (corrected)	**2793.52**	**82**				

The homogeneity-of-regression-slopes assumption, tested by including a Group × Pre-test interaction term, was satisfied, F(1, 79) = 1.06, p = 0.307, supporting the validity of the ANCOVA model. After statistically controlling for pre-test achievement, the adjusted post-test means were 22.41 for the experimental group and 21.17 for the control group. The adjusted difference between groups was not statistically significant, F(1, 80) = 1.28, p = .261, partial η^2 = .016 (H04 retained). The pre-test covariate itself was a strong and highly significant predictor of post-test performance, F(1, 80) = 37.37, p < .001, confirming that prior achievement accounted for a substantial share of post-test variance and needed to be controlled.

8. Discussion

The pattern of results across Tables 3–6 tells a coherent and methodologically important story. On raw scores alone, the experimental group appeared to outperform the control group at post-test (Table 4), and this difference was statistically significant with a moderate effect size. However, an identical-magnitude difference was already present at pre-test (Table 3), before the AI-assisted intervention began. When the amount of improvement (gain scores) is compared directly (Table 5), the two groups are statistically indistinguishable, and when pre-test achievement is formally controlled through ANCOVA (Table 6), the group difference in post-test performance is no longer significant. Together, these analyses indicate that the raw post-test advantage of the experimental group is attributable to the groups' pre-existing difference in achievement rather than to a demonstrable independent effect of the AI-assisted intervention itself.

This does not mean the AI-assisted approach was ineffective. Both groups improved substantially and significantly from pre-test to post-test, with within-group effect sizes (dz = 3.15 for the experimental group; dz = 1.44 for the control group) that are large by conventional standards. The finding is best read as evidence that AI-assisted instruction produced learning gains comparable to, rather than demonstrably greater than, regular classroom instruction over the eight-week period studied. This is consistent with meta-analytic benchmarks in the literature: VanLehn (2011) and Ma et al. (2014) report average intelligent-tutoring effects in the small-to-moderate range (d/g ≈ 0.4–0.75) relative to conventional instruction and note that effects tend to strengthen with longer exposure.

An eight-week implementation sits toward the shorter end of durations associated with detectable effects in that literature, which may partly explain why no independent advantage emerged once baseline differences were controlled. The baseline inequality itself is a well-documented risk of non-randomized, intact-group designs (Dimitrov & Rumrill, 2003) and is consistent with the caution raised by Khazanchi et al. (2025) regarding causal interpretation of quasi-experimental AI-in-education studies. Because classes, rather than individual students, were assigned to condition, any pre-existing difference between the two classes in prior teaching, classroom composition, or unmeasured factors would manifest as a pre-test difference that a raw post-test comparison cannot distinguish from a treatment effect. The ANCOVA approach adopted here, following the methodological guidance of Dimitrov and Rumrill (2003), is the appropriate way to address this, and its result a non-significant adjusted difference should be treated as the primary finding of the study, with the raw post-test comparison retained for transparency and comparison with studies that do not control for baseline differences.

The divergence between the significant raw post-test t-test and the non-significant Mann–Whitney U test on the same post-test scores (Table 4) reinforces this cautious interpretation: the significance of the raw parametric comparison appears sensitive to the specific test used and to the heterogeneity of variance detected by Levene's test, rather than reflecting a robust, assumption-independent group difference.

9. Educational Implications

- AI-assisted instruction produced mathematics achievement gains at least comparable to regular instruction over an eight-week period, supporting its viability as a supplementary instructional tool rather than a wholesale replacement for classroom teaching.
- Schools and teachers considering AI-assisted platforms should not expect large, guaranteed gains over conventional instruction within a short implementation window; realistic expectations and longer implementation periods are warranted, consistent with duration-dependent effects reported in the tutoring-systems literature (Ma et al., 2014).
- Where AI tools are piloted in some classes but not others, evaluators should anticipate and plan for baseline non-equivalence between intact groups and should analyse gain scores or use covariate-adjusted methods (e.g., ANCOVA) rather than relying on raw post-test comparisons alone.
- Consistent with recommendations from systematic reviews of AI in mathematics education (Mohamed et al., 2022), AI tools are best positioned as a complement to, rather than a substitute for, teacher-led instruction.

10. Limitations

- Non-random assignment of intact classroom groups to condition introduced baseline inequality in prior achievement, limiting causal inference despite statistical adjustment.
- The sample was drawn from a single school, limiting generalizability to other schools, grades, or regions.
- The Achievement Test in Mathematics was researcher-developed; reliability and validity evidence for the instrument should be reported to strengthen confidence in the measure. [Author: insert Cronbach's α / validity evidence.]
- The intervention period (just over eight weeks) is relatively short relative to the durations associated with larger effects in the intelligent-tutoring literature.
- The post-test distribution showed heterogeneity of variance across groups, and one group's pre-test distribution departed from normality; while robustness checks (Welch's correction, Mann–Whitney U) were applied, results should still be interpreted with appropriate caution.

12. Conclusion

This study examined the effect of AI-assisted teaching on mathematics achievement among secondary school students using a quasi-experimental pre-test–post-test control group design. Both the AI-assisted and conventionally taught groups showed substantial and statistically significant improvement in achievement over the study period, indicating that meaningful learning occurred under both conditions. While the experimental group scored significantly higher than the control group on raw post-test scores, this difference mirrored a pre-existing baseline difference between the groups, and it was not statistically significant once pre-test achievement was controlled through ANCOVA. The available evidence therefore does not establish an independent causal effect of the AI-assisted intervention over conventional instruction in this sample, although it is consistent with AI-assisted teaching being at least as effective as regular instruction over an eight-week period. These findings underscore the importance of accounting for baseline group differences in school-based quasi-experimental research and point to the need for larger, longer, and more rigorously controlled studies before firm claims about the causal effectiveness of AI-assisted mathematics instruction can be made.

References

Dimitrov, D. M., & Rumrill, P. D., Jr. (2003). Pretest-posttest designs and measurement of change. Work, 20(2), 159–165.

Khazanchi, R., Di Mitri, D., & Drachsler, H. (2025). The effect of AI-based systems on mathematics achievement in rural context: A quantitative study. Journal of Computer Assisted Learning, 41(1), e13098.

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

Mohamed, M. Z. B., Hidayat, R., Suhaizi, N. N. B., Sabri, N. B. M., Mahmud, M. K. H. B., & Baharuddin, S. N. B. (2022). Artificial intelligence in mathematics education: A systematic literature review. International Electronic Journal of Mathematics Education, 17(3), em0694.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education — where are the educators? International Journal of Educational Technology in Higher Education, 16, Article 39.