Raising the Floor, Holding the Bar:

Distributional Effects of AI-Assisted Course Design in Two Community College Mathematics Courses

Safaa Dabagh

Department of Mathematics, Santa Monica College, California, United States

dabagh_safaa@smc.edu | ORCID: 0009-0006-7752-3354

Abstract

Instructors increasingly use generative artificial intelligence (AI) to design course materials, but evidence on student outcomes is scarce and rarely looks beyond differences in means. We analyse official gradebook records from four sections of two mathematics courses taught by the same instructor at an open-access two-year college in California: introductory statistics and college algebra, each observed in a conventionally designed term and an AI-assisted term ($N = 106$). With family-wise error controlled across all ten component comparisons (Holm procedure), two results survive correction. In algebra, where AI-assisted design targeted scaffolding and preparation, final-examination scores rose 16.9 percentage points ($p = .004$, Hedges $g = 0.75$)—and quantile decomposition shows the gain is a *floor effect*: the 10th percentile rose 36 points and the share of scores below 50% fell from 31% to 5%, while the 90th percentile was essentially unchanged. In statistics, where the redesign raised assessment demand, performance was maintained on a harder instrument. Beneath both results, unproctored scores inflated toward the ceiling, and a within-design contrast shows the consequence depends on alignment: quizzes AI-designed alongside the examinations retained their value as predictors of final-examination performance, whereas platform item-bank quizzes inside an otherwise redesigned course lost it entirely. Implications for assessment design, equity, and the study of AI-assisted instruction are discussed.

Keywords: generative AI; instructional design; distributional effects; formative assessment; constructive alignment; community college; equity; academic integrity

1 Introduction

Generative AI is entering tertiary classrooms from two directions at once. Instructors use large language models to draft and refine instructional materials; students use the same tools—sanctioned or not—to complete coursework. The research literature has so far divided along the same line: studies of instructor-side AI concentrate on the properties of the materials produced (Kasneci et al., 2023; Mollick and Mollick, 2023), while studies of student-side AI concentrate on academic integrity (Cotton et al., 2024; Newton and Essex, 2024). The literature is largely silent on the question practitioners ask first: *when an instructor redesigns a course with AI assistance, what happens to student outcomes?*

Where outcome evidence exists, it is almost always reported as a difference in means. Means, however, can conceal precisely the information that open-access institutions most need: *which* students moved, and *where* in the score distribution the change occurred. A 17-point mean gain produced by lifting the weakest students out of failure has different causes, different policy value, and different design implications than the same gain produced by pushing strong students higher. Distributional analysis has a long tradition in economics (Koenker and Bassett, 1978) and is occasionally urged in education research; it is nearly absent from the emerging literature on AI-assisted instruction.

This study addresses both gaps in an unusually controlled naturalistic setting: one instructor, one institution, two different mathematics courses, each observed in a conventionally designed term and an AI-assisted term. The two-course structure permits a test of whether "the effect of AI-assisted course design" is a single quantity at all, or whether it inherits the direction of the pedagogical intention it serves: in introductory statistics the redesign aimed to raise the cognitive demand of assessment, while in college algebra it aimed to scaffold procedural mastery. A companion analysis of the same four courses' examination instruments (Dabagh, 2026) documents this divergence at the level of assessment design; the present study asks whether—and *how*—student outcomes follow.

The study makes four contributions. First, it provides outcome evidence on instructor-side AI use from complete official gradebook records rather than self-report. Second, it reports effects distributionally, identifying a floor effect that mean-only reporting would conceal. Third, it exploits a within-design contrast—one AI term whose quizzes were AI-designed in alignment with the examinations, one whose quizzes remained platform item-bank instruments—to isolate *alignment* as a moderator of whether formative assessment retains its diagnostic value in AI-rich courses. Fourth, it demonstrates a reporting discipline (family-wise error control across all component comparisons; per-course estimates before pooling) that the emerging literature on AI-assisted instruction will need as it matures.

We address four research questions:

- **RQ1.** Within each course, did performance on the proctored final examination differ between the conventional and AI-assisted terms, and do any differences survive family-wise error correction across all component comparisons?
- **RQ2.** Where in the score distribution did any change occur—at the floor, the ceiling, or uniformly?
- **RQ3.** Did the relationship between unproctored (formative) and proctored (summative) performance change between conditions, and does assessment alignment moderate that change?
- **RQ4.** Did completion of the final examination differ between conditions, and is a pooled cross-course estimate of the AI association meaningful?

2 Background

2.1 AI as an instructor's design tool

Systematic reviews of AI in higher education find the literature dominated by student-facing applications—tutoring, feedback, writing support—with instructor-facing design applications underrepresented (Zawacki-Richter et al., 2019; Lo, 2023). Position papers and practitioner accounts argue that large language models can act as a force multiplier for instructors, drafting items, examples, and scaffolds that the instructor curates (Mollick and Mollick, 2023; Kasneci et al., 2023), and discipline-specific communities have begun documenting both the capabilities and the risks (Denny et al., 2024). What this literature lacks is outcome evidence: studies connecting instructor-side AI use to what students subsequently do on assessments the AI did not grade and the student could not outsource—that is, proctored examinations.

2.2 Alignment and the function of formative assessment

Two older literatures predict where any effect of AI-assisted design should appear. Constructive alignment theory holds that learning is driven by the coherence among intended outcomes, learning activities, and assessment (Biggs, 1996; Biggs and Tang, 2011); misaligned components send students conflicting signals about what to practise. The formative-assessment literature holds that low-stakes assessment improves learning only insofar as it generates valid information about readiness—for the student and the instructor (Black and Wiliam, 1998). Retrieval-practice research adds that the practice must engage the same knowledge the summative assessment demands (Roediger and Karpicke, 2006), and cognitive-load theory explains why scaffolded materials particularly benefit novices (Sweller, 1988). Together these literatures imply a specific, testable proposition: AI-assisted design should improve outcomes *through* coherence—and formative instruments left outside the design loop should lose, not merely fail to gain, diagnostic value. Producing a fully aligned set of course materials has always been pedagogically desirable and practically unaffordable for a single instructor; one-to-one alignment of every layer is a labour problem as much as a design problem (Bloom, 1984; VanLehn, 2011). AI assistance changes the labour economics.

2.3 The open-access context

The study site is a community college: an open-access public institution in the United States that admits all applicants and serves a student population of exceptional diversity in age, preparation, and goals (Bailey et al., 2015). California's placement reforms (Assembly Bills 705 and 1705) moved students directly into transfer-level mathematics with concurrent (corequisite) support rather than prerequisite remediation, widening the preparation range within a single classroom (Mejia et al., 2020; Logue et al., 2016). In this context the lower tail of the score distribution is not a statistical abstraction; it is the policy-relevant population. Assessment-design frameworks for the resulting classrooms emphasise reasoning over recall in statistics (Carver et al., 2016; Wild and Pfannkuch,

1999; Chance, 2002; Garfield and Ben-Zvi, 2008) and universally designed, scaffolded materials across courses (CAST, 2018). Whether AI-assisted design helps precisely the students these reforms brought into the classroom is, to our knowledge, untested.

2.4 The unproctored–proctored gap

Finally, any study of AI-era coursework must reckon with the integrity environment: unproctored online work is now completable by AI tools, and reviews suggest substantial and rising rates of outside assistance in unproctored assessment (Newton and Essex, 2024; Cotton et al., 2024). This study treats the resulting score inflation not primarily as a moral problem but as a measurement problem: inflated formative scores stop carrying information about readiness, with consequences for students' self-assessment and instructors' early-warning systems.

3 Method

3.1 Design and setting

The study uses a non-equivalent-groups quasi-experimental design crossing two courses with two assessment-design conditions at a large urban California community college (approximately 30,000 students, the majority from historically underrepresented groups). The four intact sections, all taught by the same instructor with concurrent corequisite support sections, are: introductory statistics (STAT C1000, six-week intensive winter session; 2025 conventional, 2026 AI-assisted) and college algebra for students intending science, technology, engineering, and mathematics (STEM) study (MATH 4, sixteen-week semester; 2025 conventional, 2026 AI-assisted). Cohorts were not randomly assigned; all inference is associational.

3.2 Participants, data, and ethics

The analytic sample comprises the 106 students in the cleaned institutional gradebook exports: statistics $n = 28$ (2025) and $n = 27$ (2026); algebra $n = 32$ (2025) and $n = 19$ (2026). Section enrolments (29, 27, 34, and 26, respectively) serve as denominators for the completion and pass-rate outcomes reported in Section 4.5. The gradebook exports (June 2026) consolidate every graded category, including extra credit, exactly as grades were experienced by students; records were de-identified with sequential study codes prior to analysis, and no names, identification numbers, or other personally identifying information were retained in the analysis dataset. At the time of the study the institution was not conducting full institutional review board reviews; in accordance with written guidance from its Office of Institutional Research, the study proceeded as educational research with the documented approval of the department chair (March 2026). All data are de-identified records of normal educational practice from the author's own course sections, analysed and reported at the group level, with no experimental manipulation of students and no sensitive personal information.

3.3 Intervention

In the AI-assisted terms the instructor used a large language model (Claude, Anthropic) in a structured design workflow to create the required course materials used by every enrolled student. The scope of AI design differed between the two AI terms in one consequential respect. In statistics—the instructor's first AI-assisted term—AI assistance was directed at the summative and instructional layer: examinations, lecture materials, formula sheets, and study guides, with the design goal of raising the cognitive demand of assessment items (the share of higher-order items on the final examination rose from 18% to 48%; Dabagh, 2026). Both the unit examinations and the final examination were fully AI-designed under this workflow at elevated cognitive demand, with every design decision reviewed and approved by the instructor. Quizzes, however, continued to be drawn from the item banks of MyOpenMath—an open online homework platform offering publisher-style question banks, widely used in United States mathematics courses—and were not AI-designed, leaving them outside the alignment loop of the redesigned examinations. In algebra—the second AI-assisted term, by which time the instructor's workflow had matured—AI design extended to the formative layer: quizzes were AI-designed in explicit alignment with the unit and final examinations, alongside per-item worked-step scaffolding, distractors encoding documented misconceptions, a formula sheet, parallel examination forms, and an aligned preparation package (study guide, practice final with key).

A further component of the AI-assisted workflow, common to both AI terms, was *ongoing adaptation*: the instructor revised examples, scaffolds, and review materials during the term in response to the observed needs of the students enrolled in that section—a degree of mid-course responsiveness that is prohibitively time-consuming for a single instructor working manually. This difference in formative-layer treatment creates an informative within-design contrast between AI-designed-and-aligned quizzes (algebra) and platform item-bank quizzes inside an otherwise AI-redesigned course (statistics), examined in Section 4.4. Students in the AI-assisted terms additionally had optional access to an instructor-authored open-educational-resources text and companion website; uptake was not tracked, and these optional resources are not treated as part of the intervention. The intervention describes instructor-side AI use; student AI use was neither sanctioned nor measured. Table 1 summarises the structural characteristics of the four final examinations, the study's primary outcome instruments.

Table 1: Structural characteristics of the four final examinations.

	Statistics		Algebra	
Feature	Conventional	AI-assisted	Conventional	AI-assisted
Items	41 (34 MC, 7 fill-in)	25 MC	20 free-response	20 (19 MC, 1 proof)
Total points	185	100	100	100
Reference materials	Own notes	Formula sheet	Own notes	Formula sheet
Parallel forms	No	No	No	Yes (A/B)
Worked-step scaffolds	No	No	No	Yes (per item)
Bonus points	None	None	None	None

3.4 Measures

All component scores (homework, quizzes, unit examinations, final examination) and overall course percentages were taken directly from the official learning-management-system gradebook category scores, which apply the instructor's grading policies (drop-lowest rules, bonus points, extra credit) as students experienced them. Algebra attendance was computed as the percentage of attendance-tracked sessions with positive credit; in the statistics exports attendance exists only as a graded category incorporating bonuses and penalties and is not analysed. Unit-exam and overall percentages may marginally exceed 100% owing to bonus points; *the final examination carried no bonus in any term*, making it the cleanest instrument in the gradebook. Completion was defined as sitting the final examination, and students who sat a given examination are referred to throughout as its *takers*. Two derived quantities are defined here: the *ceiling rate* of a component is the share of takers scoring at or above 95% on it, and the *unproctored–final gap* is the mean of homework and quiz percentages minus the final-examination percentage.

3.5 Statistical analysis

The proctored final examination was designated the primary outcome for each course on measurement-quality grounds: it is the only instrument in the gradebook that is simultaneously proctored, bonus-free, and taken by every completer. All other components are secondary. Distributional assumptions were checked per cell: final-examination scores departed from normality in two of four cells (Shapiro–Wilk $p = .017$, algebra-conventional; $p = .014$, statistics-AI), with unequal variances across algebra conditions (Levene $p = .009$); Welch's unequal-variance t-test and the Mann–Whitney U test are therefore reported together, with Hedges g and 95% confidence intervals. Family-wise error across the ten component comparisons was controlled with the Holm procedure (Holm, 1979). Distributional location of effects was examined by quantile decomposition (10th–90th percentiles), with percentile-bootstrap 95% confidence intervals (10,000 resamples) for quantile differences and for the change in the share of scores below 50%. The sensitivity of taker-based distributional comparisons to differences in examination-taking rates was assessed with a trimming bound in the spirit of Lee (2009), comparing AI-term takers against the top equivalent fraction of conventional-term takers. Formative–summative relationships use Spearman rank correlations, which are robust to the ceiling compression documented below, with percentile-bootstrap 95% confidence intervals; the contrast between two independent correlations uses the Fisher z transformation. For algebra, analysis of covariance (ANCOVA) adjusted the condition contrast for attendance. Completion and pass rates (all enrolled students in the denominator) were compared with Fisher's exact test; because institutional conventions vary in whether a D-range or C-range grade constitutes a transfer-relevant pass, pass rates are reported at both the $\geq 60\%$ and $\geq 70\%$ thresholds. The cross-course structure was examined with an ordinary-least-squares model of final-examination score on course, condition, and their interaction. Analyses were conducted in Python (NumPy 2.2, SciPy 1.17).

4 Results

4.1 Primary outcomes and the family-wise ledger (RQ1)

Table 2 reports all ten component comparisons among final-examination takers; Figure 1 displays the four cells. Under Holm correction across the full family, exactly two results survive: the algebra final examination (+16.9, adjusted $p = .036$) and statistics homework (+14.4, adjusted $p = .010$). The primary outcome thus passes family-wise correction in algebra and shows no deficit in statistics.

Table 2: Component comparisons among final-examination takers (AI-assisted − conventional). Holm correction applied across all ten tests; † marks results surviving at $\alpha = .05$.

Course	Variable	Δ	95% CI	Welch p	Holm p	MWU p	g
Statistics (28 vs. 27)	Homework†	+14.4	[+6.2, +22.6]	.001	.010	.001	+0.93
	Quizzes	+7.1	[+0.3, +14.0]	.041	.246	.005	+0.56
	Unit exams	−9.9	[−18.1, −1.6]	.020	.140	.097	−0.66
	Final exam	+3.6	[−6.0, +13.2]	.456	>.99	.259	+0.20
	Overall grade	−0.2	[−6.6, +6.2]	.956	>.99	.755	−0.01
Algebra (32 vs. 19)	Homework	+5.5	[−11.7, +22.7]	.522	>.99	.140	+0.19
	Quizzes	+20.4	[+5.0, +35.9]	.011	.088	.064	+0.66
	Unit exams	+5.7	[−5.0, +16.4]	.288	>.99	.121	+0.33
	Final exam†	+16.9	[+5.5, +28.3]	.004	.036	.025	+0.75
	Overall grade	+8.4	[−0.8, +17.7]	.073	.365	.096	+0.52

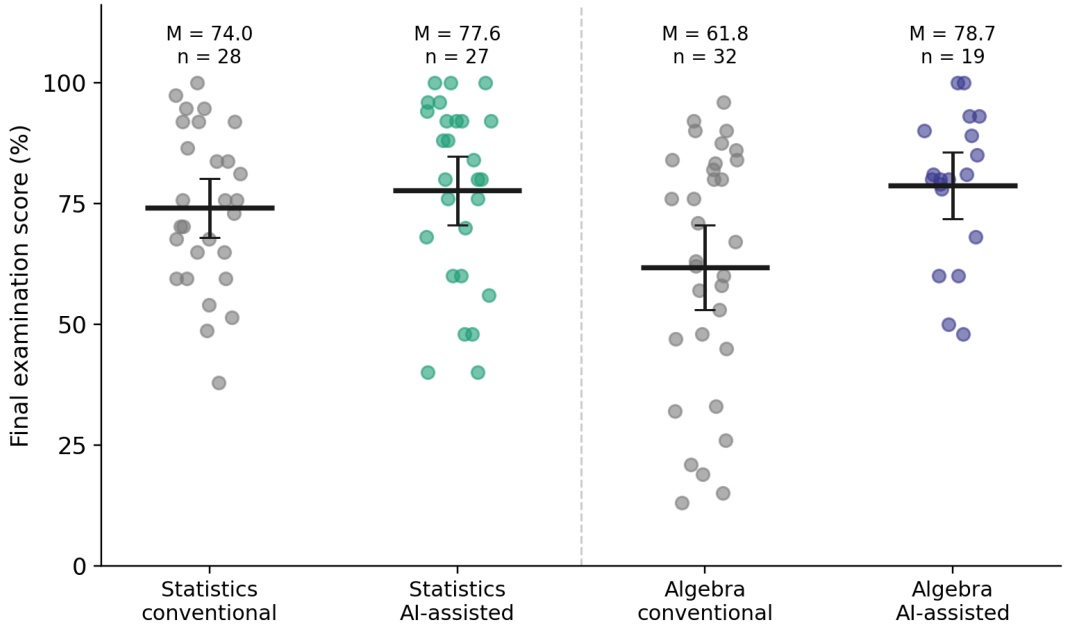

Figure 1: Final-examination scores in all four sections. Dots are individual students; horizontal bars are means with 95% confidence intervals.

Algebra. Among completers, final-examination scores rose 16.9 points (95% CI [+5.5, +28.3],

Welch $p = .004$, Mann–Whitney $p = .025$, $g = 0.75$), with variance halved (SD 25.3 to 15.2; Levene $p = .009$). Quiz scores were higher (+20.4, $p = .011$, not surviving Holm) and the overall grade directionally higher (+8.4, $p = .073$); homework and attendance did not differ, indicating comparable engagement inputs. ANCOVA adjusting for attendance left the effect essentially unchanged (+17.4, $SE = 6.4$, $p = .009$; homogeneity of slopes $p = .57$), and the advantage appeared in every attendance tercile (lowest +24.8; middle +16.7; highest +16.3).

Statistics. Final-examination performance was maintained (74.0% vs. 77.6%, $p = .46$, $g = 0.20$) even though the AI-assisted instrument was substantially more cognitively demanding (Dabagh, 2026). Homework rose (+14.4, surviving Holm) and quizzes rose (+7.1, not surviving); unit-examination scores declined (−9.9, $p = .020$, not surviving) with variance more than doubling (SD 8.1 to 19.5; Levene $p = .005$). Because the unit examinations were redesigned under the same elevated-demand AI workflow as the final (Section 3.3), this pattern is consistent with more demanding instruments discriminating more sharply rather than with cohort decline—a reading reinforced by the unchanged overall grades and pass rates on the same students. Overall grades (−0.2, $p = .96$) and pass rates (at $\geq 60\%$: 93.1% vs. 92.6%, $p = 1.0$; at $\geq 70\%$: 93.1% vs. 85.2%, $p = .41$) were statistically unchanged.

4.2 Where the change occurred: a floor effect (RQ2)

Quantile decomposition (Figure 2) shows that the algebra gain is concentrated almost entirely in the lower tail. The 10th percentile of final-examination scores rose from 22 to 58 (+36 points; bootstrap 95% CI $[+11, +57]$), the 25th from 46 to 73 (+27; CI $[+2, +52]$), and the median from 65 to 80 (+15; CI $[0, +30]$), while the 90th percentile moved from 90 to 94 (+4; CI $[-3, +16]$). The share of takers scoring below 50% fell from 31% to 5% (−26 percentage points; CI $[-44, -7]$). The interval pattern itself supports the floor characterisation: the lower-tail intervals exclude zero while the upper-tail interval does not. The scaffolding-oriented redesign did not make strong students stronger; it nearly eliminated catastrophic performance, compressing the distribution from below. In statistics, by contrast, the lower tail widened slightly (10th percentile 53 to 48) while the upper tail extended (90th percentile 95 to 98)—mild dispersion consistent with a more demanding instrument discriminating more sharply, not with uniform improvement or decline.

Because examination-taking rates differed between the algebra terms (94% vs. 73%), a natural question is whether the floor effect could reflect differences in the composition of takers rather than improved performance. Four observations argue against a compositional account. First, the all-enrolled pass rate was essentially unchanged (Section 4.5), which is difficult to reconcile with a strongly selected taker group. Second, the attendance-adjusted ANCOVA leaves the effect intact, and the advantage appears in every attendance tercile, including the lowest. Third, taker engagement inputs (homework, attendance) did not differ between terms. Fourth, and most directly, a trimming bound constructed under the extreme assumption that examination-taking is perfectly rank-correlated with performance—comparing the AI-term takers against the top-performing conventional-term takers trimmed to the AI term's 73% examination-taking rate

($k = 25$)—attenuates but does not reverse the contrast: the AI-term mean remains higher (78.7 vs. 72.7), the 10th percentile higher (58 vs. 50), the 25th percentile higher (73 vs. 60), and the sub-50% share lower (5% vs. 12%). The floor effect survives the harshest compositional assumption available to the design.

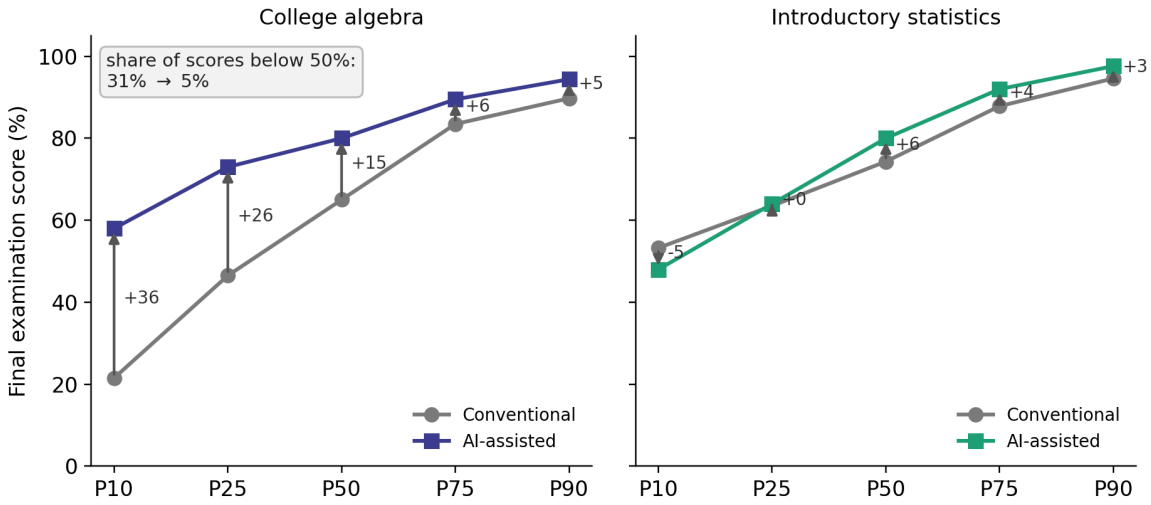

Figure 2: Quantile decomposition of final-examination scores (P10–P90), conventional versus AI-assisted term. The algebra gain is a floor effect: large movement at the 10th and 25th percentiles, minimal movement at the 90th.

4.3 Formative–summative decoupling (RQ3)

In both courses the AI term weakened the link between unproctored and proctored performance. Homework ceiling rates rose from 32% to 85% of takers in statistics, and from 22% to 63% in algebra. In statistics, the Spearman correlation between quiz average and final-examination score fell from $\rho = .25$ to $\rho = .05$ (Figure 3): in the AI term, quiz scores carried essentially no information about final-examination performance, while the proctored unit examinations retained a strong association with the final ($\rho = .81$). The unproctored–final gap among statistics takers widened from +9.6 to +17.0 points. Two mechanisms are observationally equivalent here: AI-designed formative work may be more completable by design (scaffolded, retry-friendly), and students may complete unproctored work with AI assistance. Both compress formative scores against the ceiling and drain their diagnostic value; the data cannot apportion the mix, but the consequence for practice is the same and is taken up in the Discussion.

4.4 Aligned versus platform-bank quizzes: a within-design contrast

The two AI-assisted terms differ in whether the formative layer was brought inside the AI design loop (Section 3.3): algebra quizzes were AI-designed in explicit alignment with the examinations, while statistics quizzes remained platform item-bank instruments inside an otherwise AI-redesigned

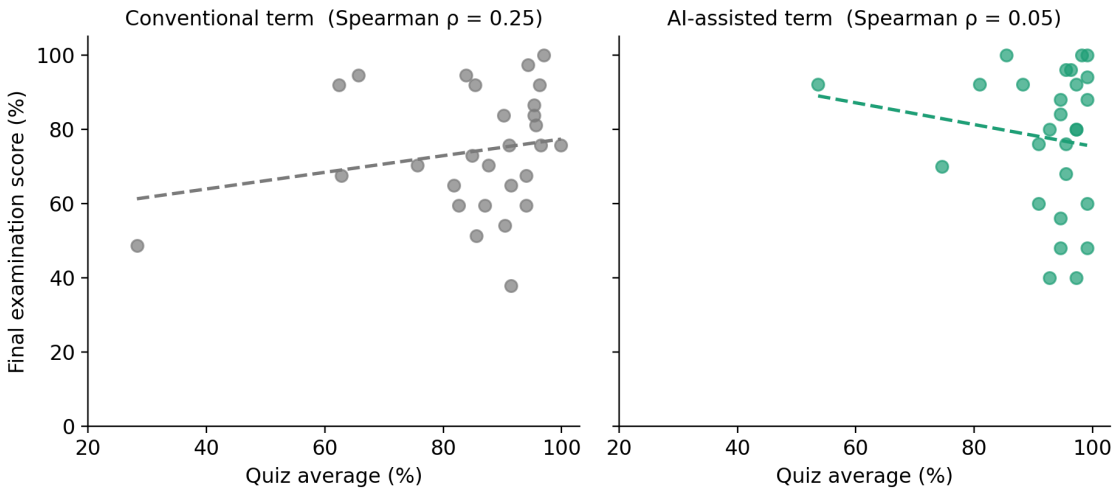

Figure 3: Quiz–final decoupling in statistics. In the conventional term quiz scores weakly track the final examination ($\rho = .25$); in the AI-assisted term the association disappears ($\rho = .05$), with quiz scores compressed against the ceiling. Dashed trend lines are illustrative; reported statistics are rank correlations.

course. Alignment theory predicts where formative signal should survive, and the data follow the prediction on every measure. In the statistics AI term, the quiz–final correlation was $\rho = .05$ (bootstrap 95% CI $[-.35, +.45]$) and the quiz–unit-exam correlation $\rho = .17$; in the algebra AI term the same correlations were $\rho = .58$ (CI $[+.11, +.86]$) and $\rho = .75$ (Fisher z for the quiz–final contrast between AI terms: $z = 1.88$, $p = .060$, reported descriptively given the cell sizes). The algebra interval excludes zero; the statistics interval is centred on it. The within-course direction of change is equally telling: relative to its own conventional baseline, the statistics quiz–final association *fell* (.25 to .05) while the algebra association *rose* (.46 to .58). Critically, ceiling inflation occurred in both AI terms (quiz ceiling rates 52% of statistics takers and 42% of algebra takers, versus 25% and 38% in the conventional terms): alignment did not prevent formative scores from inflating, but the aligned instruments retained their rank-order diagnostic value despite the inflation, whereas the unaligned platform-bank instruments lost it entirely. The decoupling of Section 4.3 is therefore not an undifferentiated property of AI-rich courses; within this design it concentrated precisely where formative instruments were left outside the AI alignment loop.

4.5 Completion and the pooled estimate (RQ4)

Completion of the final examination was near-universal in statistics in both terms (28/29 and 27/27). In algebra, completion fell from 94% (32/34) to 73% (19/26) in the AI term (Fisher exact $p = .032$). The all-enrolled pass rate was essentially unchanged at both thresholds ($\geq 60\%$: 67.6% vs. 69.2%, $p = 1.0$; $\geq 70\%$: 52.9% vs. 57.7%, $p = .80$): completer gains and completion differences offset at the cohort level, and the conclusion does not depend on the pass threshold chosen. The course $\times$ condition interaction on the final examination was +13.3 points ($SE = 7.9$, $p = .095$; underpowered, reported descriptively), and a pooled course-adjusted model yields +9.8 points (95%

CI $[+1.9, +17.7]$, $p = .016$)—a defensible summary that nonetheless averages a +3.6 and a +16.9 effect arising from different interventions serving different goals. No section experienced the pooled effect; per-course estimates are primary.

5 Discussion

5.1 Amplification, located in the distribution

The clearest reading of these results is that AI-assisted course design amplified each course's pedagogical intention—and the quantile decomposition shows precisely *where*. In algebra, the intention was scaffolded procedural mastery, and the effect materialised at the floor: a 36-point rise in the 10th percentile and a collapse in sub-50% performance, with the top of the distribution essentially untouched. For an open-access institution, this is arguably the most policy-relevant shape an effect can take: the students who moved are the students whom corequisite-era placement reform brought directly into transfer-level mathematics (Mejia et al., 2020; Logue et al., 2016). The intervention's adaptive component offers a candidate mechanism for this shape. Mid-course revision of examples, scaffolds, and review materials in response to the difficulties of the students actually enrolled is, by construction, targeted at the struggling tail, while students already performing near the ceiling have little to gain from it; this is the individualisation channel that the tutoring literature has long identified as potent but unaffordable at scale (Bloom, 1984; VanLehn, 2011). The data cannot isolate this channel from the scaffolding built into the materials themselves (Sweller, 1988), but both point the same direction—toward the floor. More broadly, the adaptive component reframes course materials as *living documents*, revised continuously in response to the students enrolled, rather than fixed artifacts set before the term; the structured AI workflow is what makes that responsiveness feasible for a single instructor at semester speed. In statistics, the intention was more demanding assessment, and performance held on a harder instrument with mild dispersion—itself a positive outcome, since the a priori concern when raising cognitive demand is collapse at the bottom. "The effect of AI-assisted design" is not one number, because AI-assisted design is not one intervention: here it was two interventions with two distributional signatures.

The completion difference in algebra (94% vs. 73% of enrolled students sitting the final examination) merits comment, although the design cannot adjudicate among explanations. Candidate accounts include ordinary cohort variation at these section sizes, differences in students' standing entering the final-examination period, and term-format differences in opportunities to defer assessment; the unchanged all-enrolled pass rate indicates that, at the cohort level, completer gains and completion differences offset. For adopting instructors the practical reading is straightforward: course redesigns should be evaluated on completion behaviour alongside performance, since neither alone summarises cohort outcomes; the trimming bound of Section 4.2 shows the performance finding here does not depend on how the completion difference is interpreted.

5.2 Formative inflation as a side effect, moderated by alignment

The decoupling results unify what would otherwise look like separate anomalies, and the within-design contrast identifies their most plausible moderator. Formative assessment serves learning only insofar as it generates valid information about readiness (Black and Wiliam, 1998); the statistics AI term shows how completely that validity can drain away when unproctored instruments sit outside the design loop, even as the surrounding course improves. Three mechanisms can produce formative inflation: AI-designed formative work may be more completable by design; students may complete unproctored work with AI (Newton and Essex, 2024; Cotton et al., 2024); and formative instruments may be misaligned with redesigned summative instruments (Biggs, 1996). The contrast of Section 4.4 cannot apportion the first two, but it isolates the third: inflation occurred in both AI terms, yet only the AI-designed, examination-aligned quizzes retained rank-order diagnostic value, while the platform-bank quizzes inside the redesigned statistics course lost it entirely—consistent with retrieval-practice findings that practice benefits transfer only when it engages the knowledge the criterion task demands (Roediger and Karpicke, 2006).

The practical implication is sharper than a generic integrity warning: *when a course is redesigned with AI, formative instruments left outside the design loop stop measuring readiness.* Four design responses follow. (1) Bring the formative layer into the same structured workflow as the summative layer, so quizzes and examinations are built as one aligned object. (2) Where platform item banks must remain, restore signal independently—frequent low-stakes proctored checks, oral verification, or in-class components tied to homework—so struggling students are identified before, rather than at, the summative examination. (3) Treat formative ceiling rates and formative–summative correlations as routine course-health telemetry; both are computable from any gradebook. (4) Expect inflation pressure in all AI-era courses and design grading weight accordingly, anchoring course grades to proctored instruments.

5.3 Distributions and families, not means and singletons

Two reporting practices materially changed what this study shows, and we commend both to future work on AI-assisted instruction. First, the mean difference in algebra (+16.9) conceals the finding's actual character; quantile decomposition reveals a floor effect with direct equity relevance. Second, the Holm ledger across all ten comparisons separates the two robust results from the suggestive ones, and the pooled cross-course estimate—though significant—averages unlike effects that no section experienced. Mean-only reporting and premature pooling would have told a tidier and substantially less true story.

6 Limitations

- Non-random, non-equivalent cohorts; all estimates are associations. Course-retake status could not be linked and is a plausible confound.

- The examinations are different instruments across terms within each course; instrument difficulty cannot be fully separated from cohort performance. In statistics the AI-term instrument was more demanding and in algebra more scaffolded, so the bias plausibly runs in opposite directions across courses.
- Component analyses describe the analytic sample; the all-enrolled pass rate provides the cohort-level summary.
- Optional supplementary resources (an instructor-authored OER text and companion website) were available in the AI-assisted terms with untracked uptake; their contribution cannot be separated from that of the required components.
- Samples are small (19–32 takers per cell); confidence intervals are wide, the interaction test is underpowered, and tail quantile estimates rest on few observations. The design partially compensates by replicating the comparison across two courses within one instructor, but multi-instructor replication is required before generalisation.
- Student AI use was not measured; the decoupling mechanism cannot be apportioned between material design and student-side AI use. Process components of the intervention, including mid-course adaptation, are documented by the instructor rather than independently logged.
- The two courses differ in format (six-week intensive vs. sixteen-week semester), content, and population; the course-level contrast confounds design intention with these differences. The two AI terms additionally differ in implementation maturity: the algebra term was the instructor's second structured-AI implementation and extended AI design to the formative layer, so the alignment contrast of Section 4.4 confounds instrument alignment with instructor experience.
- Single instructor and institution; the two-course design probes generalisability across courses, not across instructors.

7 Conclusion

Across four sections of two community college mathematics courses taught by the same instructor, AI-assisted course design was associated with a large, family-wise-robust gain on the proctored algebra final examination that quantile decomposition identifies as a floor effect—a 36-point rise in the 10th percentile and near-elimination of sub-50% performance—and with maintained performance on a substantially more demanding statistics examination. Beneath both results, unproctored scores inflated toward the ceiling; a within-design contrast shows that formative instruments AI-designed in alignment with the examinations retained their diagnostic value despite the inflation, while platform item-bank instruments outside the design loop lost it entirely. The findings support treating "AI-assisted design" as a family of interventions defined by the pedagogical intentions they

amplify, reporting outcomes distributionally rather than as means alone, and bringing the formative layer into the same structured AI workflow as the summative layer so that formative assessment keeps its signal. For open-access institutions, the floor effect is the headline: the students who gained most are the students placement reform brought into the room. A pre-registered follow-up with common blueprinted final examinations, recorded retake status, measured student AI use, and multiple instructors is the natural next step, and these data provide its effect-size foundations.

Data availability

The de-identified analysis dataset (`AI_2x2_master_v2.csv`), built solely from cleaned institutional gradebook exports (June 2026), and analysis code are available to reviewers upon request and will be deposited in an open repository upon acceptance. All scores reproduce official gradebook category computations.

Declaration of interest

The author used the AI system studied here (Claude, Anthropic) to assist in drafting course materials as described in Section 3.3. Manuscript preparation was assisted by AI tools under the author's direction; the author takes full responsibility for the content.

References

Bailey, T., Jaggars, S. S., & Jenkins, D. (2015). *Redesigning America's community colleges: A clearer path to student success*. Harvard University Press.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364.

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

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

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Rowell, G. H., Velleman, P., Witmer, J., & Wood, B. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association.

CAST. (2018). *Universal design for learning guidelines version 2.2*. `http://udlguidelines.cast.org`

Chance, B. L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education, 10*(3).

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International, 61*(2), 228–239.

Dabagh, S. (2026). Examination design under AI assistance: A Bloom's taxonomy and universal design analysis of the same four course sections. Manuscript under review.

Denny, P., Prather, J., Becker, B. A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B. N., Santos, E. A., & Sarsa, S. (2024). Computing education in the era of generative AI. *Communications of the ACM, 67*(2), 56–67.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies, 76*(3), 1071–1102.

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410.

Logue, A. W., Watanabe-Rose, M., & Douglas, D. (2016). Should students assessed as needing remedial mathematics take college-level quantitative courses instead? A randomized controlled trial. *Educational Evaluation and Policy Analysis, 38*(3), 578–598.

Mejia, M. C., Rodriguez, O., & Johnson, H. (2020). *A new era of student access at California's community colleges*. Public Policy Institute of California.

Mollick, E., & Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *SSRN Working Paper 4391243*.

Newton, P. M., & Essex, K. (2024). How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review. *Journal of Academic Ethics, 22*(2), 323–343.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

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

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*(3), 223–248.

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*, 39.