ISSN: 2458-8989

Natural and Engineering Sciences

NESciences, 2026, 11 (3): 828-844 doi: 10.28978/nesciences.263065

Dynamic Visualization and Distributional Equity in Learning Double Integrals: Quasi-Experimental Evidence and Implications for Quantitative Bioscience Education

Reinaldo Antonio Guerrero-Chirinos[1]*, Marco Antonio Ayala-Chauvin[2]

[1*] Universidad Técnica Particular de Loja (UTPL), Loja, Ecuador. ORCID: https://orcid.org/0000-0003-0499-7453, E-mail: raguerrero12@utpl.edu.ec

[2] Universidad Técnica Particular de Loja (UTPL), Loja, Ecuador. ORCID: https://orcid.org/0000-0002-0084-6773, E-mail: maayala5@utpl.edu.ec

Abstract

An integral over two variables is one of the most challenging topics in multivariable calculus because it requires coordinating the algebraic characterization (the iterated integral) with a spatial view of the region of integration and the solid under the integrand. This study investigated whether a dynamically generated GeoGebra visualization that supports this coordination would affect that coordination. The respondents were 48 students in two Civil Engineering class sections enrolled in the same term (April-Aug 2026). A quasi-experimental design with a non-equivalent control group and a post-test was applied; a diagnostic test was not used in this design. The Shapiro-Wilk test indicated non-normality in 3 of 4 distributions, so the Mann-Whitney U test, Kruskal-Wallis H test with Dunn multiple comparisons, Wilcoxon signed-rank test with effect sizes (Cliff's δ), and bootstrapped CIs were used for comparisons. Improvement was significantly greater for both groups, but the magnitude differed (raw = larger for the experimental group, Mdn = 5.00 vs. 2.00, U = 404.50, p = .016, δ = 0.405; final grades were not different, Mdn = 3.00 vs. 2.00, U = 289.50, p = .982). Some findings were more robust than others - no student exposed to the applet scored lower than 4.50 or fell below the diagnostic, whereas three students in the control group did, and two in the control group regressed, while the terminal variance was 2.72 times smaller. The applet was, therefore, a tool for distributing achievement gains but not for increasing maximum achievement. The study discussed the implications for the cases of integrated rate laws, Michaelis-Menten curves, and metabolic fluxes, as the same conversion from symbolic to graphical occurs in all three cases, making these findings applicable to quantitative bioscience education, which has yet to be tested.

Keywords

Double integrals; GeoGebra; dynamic visualization; non-parametric statistics; engineering education; learning equity; quantitative bioscience education.

Available Online: 10/09/2026

1. Introduction

It is a familiar pattern in multivariable calculus exams over many years: once a student easily solves a single integral, they get stuck – almost invariably – when asked to set the limits of a double integral over a region enclosed by two curves. The trouble has nothing to do with the antiderivative algorithm and everything to do with the previous step — reading an inequality as a region of the plane and determining which variable ranges over a fixed interval and which variable is the domain's range. This challenge has been reported many times in the calculus education literature and suggests that this is a

*Corresponding Author: Reinaldo Antonio Guerrero-Chirinos, E-mail: raguerrero12@utpl.edu.ec

representational rather than a procedural barrier (Rasmussen et al., 2014). That's because the double integral requires students to navigate two semiotic worlds not typically afforded in traditional instruction. The algebraic register renders the object as an ordered sequence of symbols subject to explicit rules, or an iterated integral. The graphical register represents it as the volume under a surface over a plane area, and the student is asked to imagine it without perception. Carvalho et al. (2024) demonstrated that the concept of two-variable Riemann sums can be made concrete again by using the GeoGebra workspace as a three-dimensional visualization of the partitions. The result is similar to that found by Tatar and Zengin (2016) in the one-dimensional domain: Teachers who changed both lower and upper sums in the dynamic setting recognized the relationship between them rather than memorizing it or indexing it.

The most suitable theory to explain this change is Duval's (2006) theory of the registers of semiotic representation, which holds that understanding a mathematical object requires the ability to convert between registers. Converting one-to-one functions is particularly challenging for three-variable functions, due to students' schemata for Cartesian space (Trigueros & Martínez-Planell, 2010), the conceptualization of functions in two dimensions without the associated attribute (Martínez-Planell & Trigueros, 2012), and the use of covariational reasoning (Weber & Thompson, 2014). Ayala Chauvin and Luna Romero (2025) brought this concept to the Ecuadorian engineering classroom by designing an applet for implicit differentiation, in which learners' questions showed an increase in their geometric understanding of an object that had proved difficult in classroom conversations. Their work, similar to the current study, included population and institution but not a comparison population and different content. The algorithm does not involve any of the following decisions required to compute the integral of f(x, y) over a region R bounded by curves: recognizing R, choosing the order of integration, writing the inner limits in terms of the outer variable, and ensuring that the choice does not partition R into subregions. All decisions are based on a mental picture of the area; otherwise, students resort to memorized standard cases, which break down when they encounter a novel region. So, a technological resource alone cannot be enough to attract students to the area—students should learn how to modify this resource and how that modification is expressed within the limits of the integral.

When a static relationship is made into an observable covariation, it is this covariance (not the picture itself) that enables conversion between registers. There is substantial and unrivaled evidence of the effects of dynamic mathematics software. Chan and Leung (2014) found good results in terms of achievement from quasi-experimental studies conducted between 1990 and 2013. Ji et al. (2024) conducted a 3-level meta-analysis of 107 studies and found an intermediate effect (d = 0.632) and that the length of the interventions, their cultural setting, and their focus on autonomous software use were able to moderate this effect. When the analysis is limited to the school population, He et al. (2024) find a medium-size effect (g = 0.820) but with broad intervals and a high level of heterogeneity, while Zhang et al. (2023), in the analysis of the GeoGebra, obtain a medium to strong effect (Hedges's g = 0.653) that is greater in smaller samples and shorter interventions. The same direction is indicated by primary studies with regard to functions and limits (Zulnaidi et al., 2020), geometry with delayed retention measures (Birgin & Topuz, 2021), and sections of solids (Chivai et al., 2022).

The integration of pedagogical innovation in the development of mathematical thinking in Ecuadorian classrooms is also related to pedagogical innovation more broadly (Guerrero Chirinos et al., 2025a). Compared with its usual attention, the regularity reported by Zhang et al. (2023) should be given greater consideration. An effect that increases as sample size decreases could indicate more controlled implementation, publication bias, or statistical instability. In their study of GeoGebra research between 2010 and 2020 (Yohannes and Chen, 2021), both researchers noted the disproportionate number of papers on the positive effects of GeoGebra, with only a small number addressing the possibility that those effects might not occur. Wassie and Zergaw (2019) also cited other constraints not captured in celebratory accounts, including technological levels and student-to-computer ratios. Another lesson is that when you're talking about the effects of the software, you're usually talking about the effect of the instructional design mediated through the software.

The nine-step cycle model for teaching with GeoGebra in class was adapted from Vygotsky's model, and Bedada and Machaba (2024) restructured the teaching sequence for differential calculus concepts around it, leading to enhanced conceptual and procedural understanding among students, who also found the scaffolding provided by the model as useful as the software itself (Bedada & Machaba, 2022). Takači et al. (2015) reached the same conclusion – that the effectiveness of their GeoGebra environment depended on students' simultaneous consideration of the algebraic and graphical forms of each step, in effect creating an immediate feedback loop. Taxonomies of pedagogical opportunities also agree that how teachers design, believe, and orchestrate the use of technology determines the effects it produces for tasks, classroom interaction, and course goals (Pierce & Stacey, 2010), and recent surveys of teaching with digital technology emphasize the importance of teachers' design decisions, beliefs, and

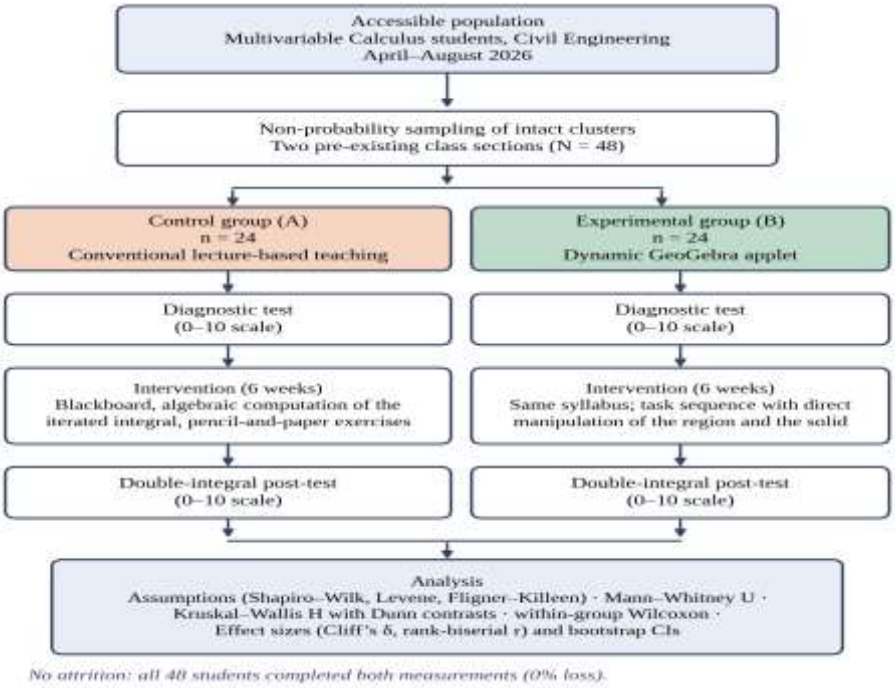
their orchestration of technology use that influence the effects of technology (Clark-Wilson et al., 2020; Gonscherowski & Rott, 2022; Thurm & Barzel, 2020, 2022). The power of GeoGebra lies in synthesizing modeling, visualization, and coding in meaningfully designed activities (Ziatdinov & Valles, 2022) and in the degree of conjecturing and visualization provided by exploratory applets (Dockendorff & Solar, 2018). There are still 2 zones of darkness. The former is thematic; most experiments on GeoGebra are related to Euclidean geometry, trigonometry, one-variable functions, and differential calculus, while experiments on the double integral are mainly suited for descriptive or design purposes and are seldom compared with another state of the problem. The latter is the latter. The second is related to the reporting of effects.

The most popular approach is to compare posttest means, after which significance is assessed, but outcomes are unclear when distributions are skewed and have a ceiling. While Takači et al. (2015) and Radović et al. (2020) found overall positive results for the experimental groups, neither examined the bottom end of the distribution, where decisions about continuing in university engineering programs are made. If most students in the class attain the maximum score, the data are truncated, and two students can have very different learning profiles with the same medians. As Laderas Huillcahuari et al. (2023) suggested in an analogous evaluation of GeoGebra in the calculus domain, rank-based statistics and dominance effects like Cliff's Δ are more suitable to this situation because they estimate the probability that a randomly chosen student in one group performs better than a randomly chosen student in the other group. Three gaps follow. First, there are no quasi-experimental comparisons among engineering students with a control group that specifically measure the effect of a dynamic applet, including one to study the double integral in Cartesian coordinates. Second, evidence often favors the mean for summarizing, even when skewness and ceiling and floor effects make it inappropriate. Finally, and most importantly, virtually no research investigates whether software effects are equally distributed across students with varied starting points. The added value of a resource that raises the average by enhancing the work of bright students is quite distinct from a resource that lifts up students who started out with gaps. Integration is important in the context of the life sciences as well as engineering, as students study enzyme kinetics and metabolic modeling, which may not often have quite as strong math backgrounds. The goal was to model the impact of a dynamic GeoGebra System on learning the double integral in the Cartesian system and to compare two groups of students without making any assumptions (i.e., using nonparametric tests). The specific objectives were to confirm the initial equivalence of the two groups, estimate within-group change and between-group differences in both final level and change in level, and characterize the shape of the resulting distributions, particularly regarding dispersion and the lower tail. Three outcome functions are defined: the final level Y, the raw gain $G = Y - X$ for some diagnostic score X, and the normalized gain $g = (Y - X)/(10 - X)$. The following hypotheses were formulated: H_{01}, that there is no difference in the diagnostic; H_{11}, that the final grades in the applet group can be characterized as higher than those of the non-applet group; H_{12}, that the raw gain in the applet group is greater than the raw gain in the non-applet group; and exploratory, H_{13}, that the dispersion of the grades after application of the applet is lower than the dispersion of the grades after application of the non-applet. To ensure the reader is aware of the evidence, the data reported were used to support H_{12} and were used descriptively to support H_{13}; the discussion is framed according to the imbalance between support for the two hypotheses, which we consider the chief contribution of the study.

2. Materials and Methods

This study used a non-equivalent control group, diagnostic measurement, and a post-test design within a quantitative, explanatory approach. Because students were assigned to sections before the study was constructed, there was no individual randomization, and the groups cannot be considered independent units for inference. They must be treated as whole sections; this prevents legally derived inferences and precludes the assumption of initial equivalence, which must be tested empirically. The research team randomly assigned the experimental sections before the diagnostic was completed, without knowing which group had previously experienced the instructional condition on the diagnostic. The teacher and the students interacted with the conventional part of the lesson, and the dynamic applet was used (Figure 1).

Figure 1. Flow diagram of the quasi-experimental design



This sample consisted of two groups of 24 Civil Engineering students at a public-mission private university who were enrolled in Multivariable Calculus during the April-August 2016 Check-up. Sampling was non-probabilistic, purposeful, and influenced by the existing section layout. In the original data, Group A was the control group (the usual lecture-based teaching), while Group B was the experimental group (using the applet with the homeworks). No student was missing either measurement; data were recorded for all 48 students, with no attrition, missing data, or imputation of missing data. Participation was compulsory; the subject was recognized as needing to be repeated in institutional problems, and students had a very stable and reliable computer lab connection, which is not necessarily found elsewhere in the region. The independent variable was the instructional condition (IC) (conventional vs. applet-mediated). The primary dependent variable was the double-integral posttest, measured on a continuous 0-10-point scale in steps of 0.25. It was used to derive raw and normalized gains and served as a covariate and the initial equivalence criterion in the diagnostic score. When a student received the highest diagnostic score, the denominator would be zero, and analyses of this variable did not include students in these conditions because they resulted from dividing by zero (see each analysis table).

The construction of both instruments was carried out by the teaching team based on the Syllabus table of specifications. The diagnostic assessed the following prerequisites: plane regions as represented by inequalities, elementary iterated integrals, and the geometric interpretation of the single integral. The post-test assessed the target content: identifying and describing the plane region of integration, justifying the integration order, setting up integration limits and computing the integral, and interpreting the solution as a volume. Following the interrater-agreement procedure and the four-phase standardized model by Guerrero Chirinos (2025b), which gradually integrates Kendall W with qualitative review of items with lower agreement, content validity was established by rejudging the items using a procedure that includes rewriting the items and, when necessary, sending them to a second round of interrater agreement with the judges.

Grading was conducted independently by two evaluators who used a four-criterion analytic rubric, with any disagreements resolved by consensus and differences never exceeding one point. Instructors, syllabi, sessions, and assessments were shared between both groups. Both groups studied the double integral over rectangles over six weeks (two 2-hour sessions each week), the double integral over Type I and Type II regions, extending the integral where appropriate, and changing the order of integration in the double integral, as well as applications of the double integral to volumes and areas. The distinction was due to the region of integration. For the control group, the region was sketched on the blackboard, and the limits were found using analytic calculation before the practical solution using pencil and paper. In the experimental group, the session started with an applet task. The applet showed the plane region formed by two curves, upon which students could move the sliders that changed these parameters; a second plane region – the iterated integral – varying in real time as the first region was changed, and a

third plane region – the iterated solid – appeared, each with a 3D view of the solid whose volume is represented by the corresponding integrand. When moving a slider, students said in advance what they expected the limits to change to, then verified their prediction and explained any difference between their answer and the answer obtained from the applet; when asked to formulate the region given an iterated integral, they compared the region they obtained with the one generated by the applet.

The design was inspired by the principle of scaffolding embedded in the task rather than in the artefact (Bedada and Machaba, 2024) and by a mechanism for instant verification of every step in the activity in dynamic environments (Takači et al., 2015). The diagnostic was administered in the first week, with no notice and under the same conditions in both sections. The intervention lasted for the following 6 weeks, and the post-test was given in week 8 as a 90-minute face-to-face session, with both groups without devices, to observe changes due to transferred learning rather than a device advantage during testing. All inferential decisions were preceded by assumption checks. The Shapiro-Wilk test, complemented by skewness and kurtosis, was also calculated, and the Kolmogorov-Smirnov statistic on standardized scores was used as a splitting criterion for normal values. Two variants of the Levene test, both median-centered and robust to non-normality (Brown & Forsythe, 1974), were used for the test of homogeneity of variance, and compared to the rank-based Fligner–Killeen test, which performed well in comparative studies of homogeneity tests (Conover et al., 1981): the mean-based Levene and the Ansari–Bradley tests are also reported. All between-group comparisons were conducted using the Mann–Whitney U test because three distributions were not normally distributed, as previously noted, in the diagnostic, post-test, and gain results. These four conditions (diagnostic procedures and post-test measures of each group) were compared together using the Kruskal–Wallis H test corrected for ties (Kruskal & Wallis, 1952), and within groups the change from diagnostic to post-test was compared by the Wilcoxon signed-rank test (Dunn, 1964).

The Cliff's Δ, the rank-biserial correlation, and the Vargha–Delaney A statistic were used to estimate effect sizes; 95 percent bootstrapped (10,000 resamples) confidence intervals were used; and distributional shift was quantified using the Hodges–Lehmann estimator and the Moses interval. Contrasts were repeated without students at the highest end of the diagnostic score distribution to serve as robustness checks, and Quade's rank analysis of covariance was used to control for possible prior student levels, as suggested to the researcher by Quade (1967). The chi-square test was used to compare achievement levels, and the proportion correct was compared using Fisher's exact test. Analyses used Python 3.12 (SciPy 1.17, NumPy 2.4, Matplotlib) with $\alpha = .05$, two-tailed. The non-parametric statistical procedures were selected according to Laderas Huillcahi et al. (2023), who addressed a similar problem. The principal problem was derived from the ethical consideration of the power imbalance between the researcher and students who were exclusively relying on him for their education. The respect tradition of the Nuremberg Code and the Declaration of Helsinki, which holds that "the end justifies the means," was replaced by "the good intentions of the researcher do not protect anybody; the independent review is required since it is not possible to judge whether the project is appropriate in an objective way, as the researcher has a personal interest in the project but the participant does not" (Castro-Molina 2020; Brítez Silvero 2016). All 48 subjects gave written informed consent.

To reduce pressure, forms were collected by someone outside the teaching team; the principal investigator did not know who had consented until after the final grades were published; and participation did not involve any additional activity, as both tests were part of the regular assessment curriculum. Because the available research was descriptive rather than evaluative, the comparison was warranted, given genuine uncertainty at the outset about the value of the applet. After post-testing, both the experimental and the control group had access to the applet and all associated tasks for two additional sessions before the course's final examination, and both groups were presented with the results, including some that did not support the initial hypothesis. The data were anonymized prior to analysis and stored in a secure institutional repository. The plan for data analysis was determined beforehand, all contrasts were presented, and no cases were omitted. Given the ethical and didactic dilemmas raised by the use of digital tools in higher education (Sánchez Romero et al., 2026), the computational tools were employed solely for statistical computation and figure generation, as declared.

3. Results and Discussion

Results are presented in an inferential sequence: description, assumption checks, initial equivalence and effects. The four grade distributions and two calculated gain variables are summarized in Table 1.

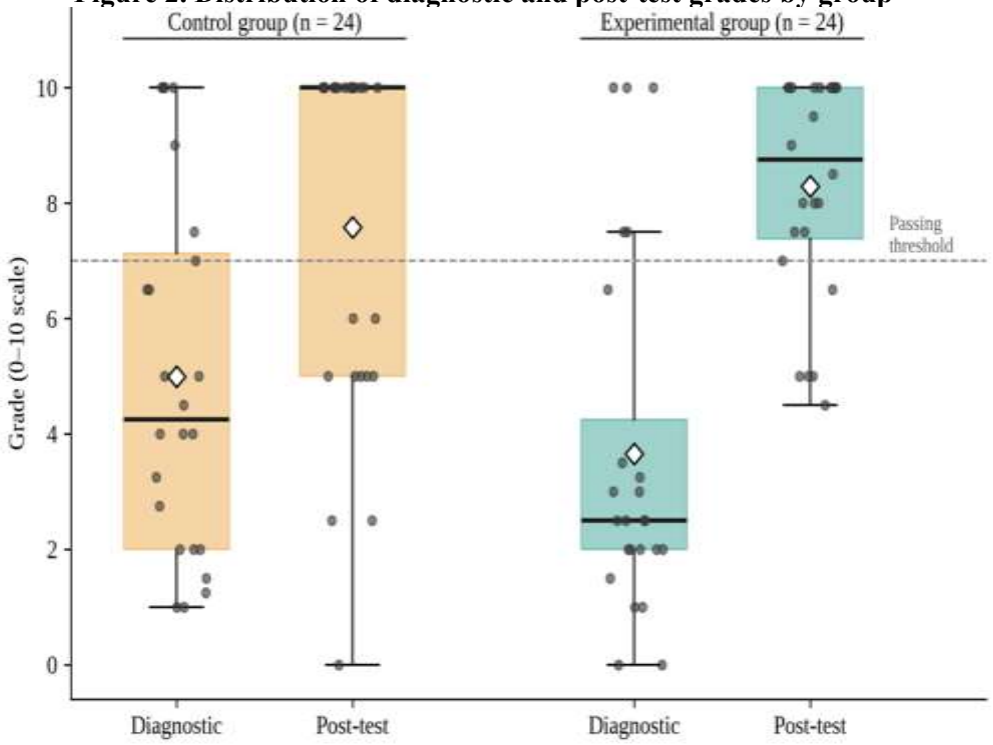
Table 1. Descriptive statistics for grades and gains by group and time point (0–10 scale)

Distribution	n	M	SD	SE	Mdn	Q1	Q3	Skewness	Kurtosis
Control – diagnostic	24	4.99	3.14	0.64	4.25	2.00	7.13	0.42	−1.14
Control – post-test	24	7.58	3.16	0.65	10.00	5.00	10.00	−0.90	−0.39
Experimental – diagnostic	24	3.66	3.12	0.64	2.50	2.00	4.25	1.14	0.07
Experimental – post-test	24	8.29	1.92	0.39	8.75	7.38	10.00	−0.80	−0.67
Gain – control	24	2.59	2.82	0.58	2.00	0.38	4.25	0.65	−0.15
Gain – experimental	24	4.64	2.72	0.55	5.00	2.50	6.81	−0.38	−0.94

Note. Own elaboration based on grade records for the April–August 2026 term. M = arithmetic mean; SD = sample standard deviation; SE = standard error of the mean; Mdn = median; Q1 and Q3 = first and third quartiles. Gain is the post-test grade minus the diagnostic grade. Skewness and kurtosis coefficients are bias-corrected.

Both groups began with low scores - a median of 4.25 and 2.50 out of 10, respectively, indicating a lack of familiarity with the content for most students. They are now better, but in different ways. The mean is 7.58 and the median is 10.00. The gap (in this case, two and a half points) indicates that this distribution is strongly left skewed, with a group of maximum scores "capturing" the median at the top and a tail of low scores "capturing" the low mean. This shows that its standard deviation (3.16) was the same as the diagnostic value (3.14): the class ended up as spread out as it started and simply moved in the upward direction. The mean was very similar to the experimental group's mean (8.29) and its median (8.75); in the experimental group, the standard deviation decreased from 3.12 to 1.92. The gap between the highest and lowest scores was greatest at the extremes, with the minimum score in the post-test in the control group being 0.00 and in the experimental group, 4.50 (Figure 2). The conventional instruction also generated the highest failure scores of 14 vs. 10 but also resulted in the highest scores for perfection.

Figure 2. Distribution of diagnostic and post-test grades by group



Note. Own elaboration. Each box spans the interquartile range; the thick horizontal line marks the median and the white diamond the arithmetic mean. Overlaid points are individual grades with random horizontal jitter to avoid overlap. The dashed line marks the passing threshold (7.00). Whiskers extend to 1.5 times the interquartile range.

Of the four grade distributions, three were not normal (Table 2). The most extreme case was the control post-test (W = 0.749): if 58.3% of observations make up the scale maximum, the distribution can no longer be considered ‘continuous' in a useful way. Gain distributions were normal; thus, the irregularity in final grade is due to ceiling truncation, and non-parametric treatment was retained for consistency. The median-centered Levene and Fligner–Killeen tests were not significant and did not reject homogeneity of variance; however, the mean-centered Levene test was significant for the post-test, indicating a discrepancy between the median- and mean-centered tests that will be examined further.

Table 2. Normality and homogeneity-of-variance tests

Distribution / comparison	S–W W	p	K–S D	p	Levene (Mdn) F	p
Control – diagnostic	0.903	.025	0.124	.813	—	—
Control – post-test	0.749	< .001	0.361	.003	—	—
Experimental – diagnostic	0.814	< .001	0.270	.049	—	—
Experimental – post-test	0.822	< .001	0.230	.133	—	—
Gain – control	0.950	.271	—	—	—	—
Gain – experimental	0.933	.116	—	—	—	—
Diagnostic: control vs. experimental	—	—	—	—	0.556	.460
Post-test: control vs. experimental	—	—	—	—	1.482	.230
Gain: control vs. experimental	—	—	—	—	0.017	.898

Note. Own elaboration. S–W = Shapiro–Wilk test; K–S = Kolmogorov–Smirnov statistic on standardized scores; Levene (Mdn) = median-centered Levene test (Brown–Forsythe variant). The complementary Fligner–Killeen test yielded $\chi^2 = 0.759$ (p = .384) for the diagnostic and $\chi^2 = 0.042$ (p = .839) for the post-test.

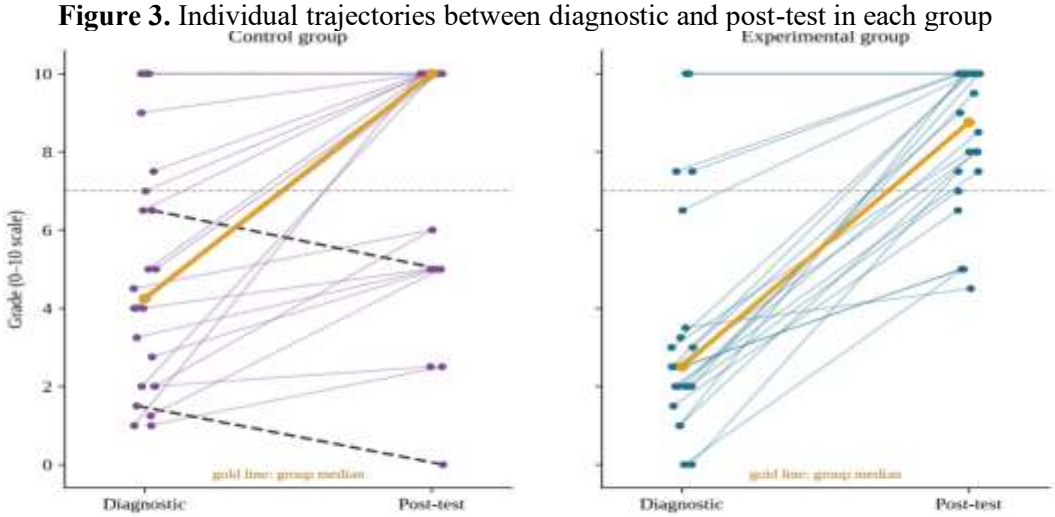
The diagnostic comparison yielded U = 213.50 (z = −1.544; p = .125; δ = −0.259, 95% CI [−0.578, 0.083]), so H_{01} was retained. Non-significance does not mean equivalence, however: the interval contains small-to-medium effects; 75% of members of the experimental group were below 4.00 points compared with 37.5% from the control group; and the median score for the experimental group was 2.50, fewer, if anything, than the median score for the control group (4.25). There were significant improvements for both groups (Table 3). Cohort negative ranks in the control group W = 11.00; and in the experimental group W = 0.00, indicating that no efficiency was lowered, and all who improved advanced on the scale, thus having a rank biserial = 1.000. This shift towards the within-group median was approximately twice as large in the experimental group (4.75 points) as in the control group (2.50 points). Individual tracks are shown in Figure 3 behind the above statistics.

Table 3. Within-group comparison between diagnostic and post-test (Wilcoxon signed-rank test)

Group	Effective n	W	z	p	r	Rank-biserial r	HL	Ties
Control	20	11.00	−3.509	< .001	0.785	0.895	2.50	4
Experimental	21	0.00	−4.015	< .001	0.876	1.000	4.75	3

Note. Own elaboration. Effective n excludes cases with no change, shown in the Ties column. W is the smaller signed-rank sum; r = |z|/√(effective n); HL = Hodges–Lehmann estimator of the within-group median shift, in points on the 0–10 scale.

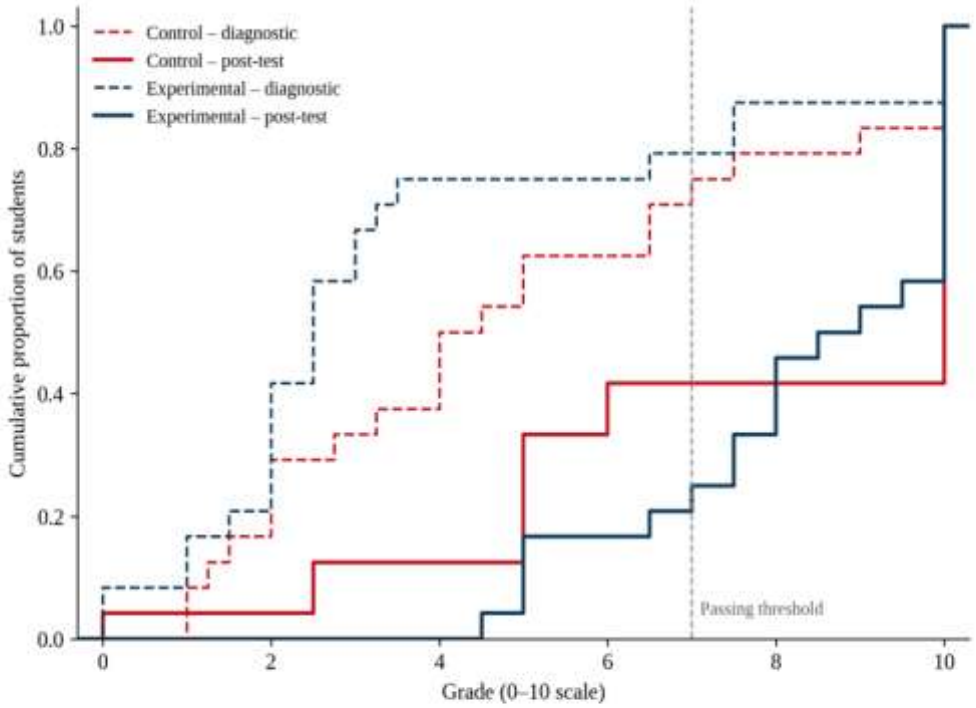
Figure 3. Individual trajectories between diagnostic and post-test in each group



Note. Own elaboration. Each line represents one student. Thin solid lines show improvements, dotted lines show no change, and thick, dark dashed lines show regression. The gold line connects the diagnostic median to the post-test median for each group. The horizontal dashed line marks the passing threshold. Minimal random horizontal jitter was applied to reduce overlap of identical scores.

The post-test comparison showed U = 289.50 (z = 0.033; p = .982; δ = 0.00 points, 95% CI [−0.313, 0.318); Vargha–Delaney A = 0.503, meaning the probability that an experimental student outperforms a randomly selected control student is as if playing a coin toss. Accordingly, H_{11} was deemed unsuitable and rejected. This seeming equivalence is actually two very differently shaped distributions (Figure 4). The outcome of the post-test ECDFs intersect at approximately 8.5 points: where the value on the graph is below 8.5, the value of the experimental curve is less than the control curve, indicating that fewer experimental students rated low on the test; whereas the value above 8.5 changes, indicating that more experimental students rated high on the test. When distribution functions cross, no measure of central position can describe their relationship – in this instance the control median (10.00) is greater than the experimental median (8.75); and the control 5th percentile (2.50) is less than the experimental (5.00). However, Mann–Whitney combines the full range, and reports a null difference due to remaining equal.

Figure 4. Empirical cumulative distribution functions of grades by group and time point



Note. Own elaboration. Each step curve shows the proportion of students whose grade is less than or equal to the value on the horizontal axis. Solid curves correspond to the post-test and dashed curves to the diagnostic. The vertical dashed line marks the passing threshold (7.00). The post-test curves cross at about 8.5 points, indicating non-uniform stochastic dominance between the groups.

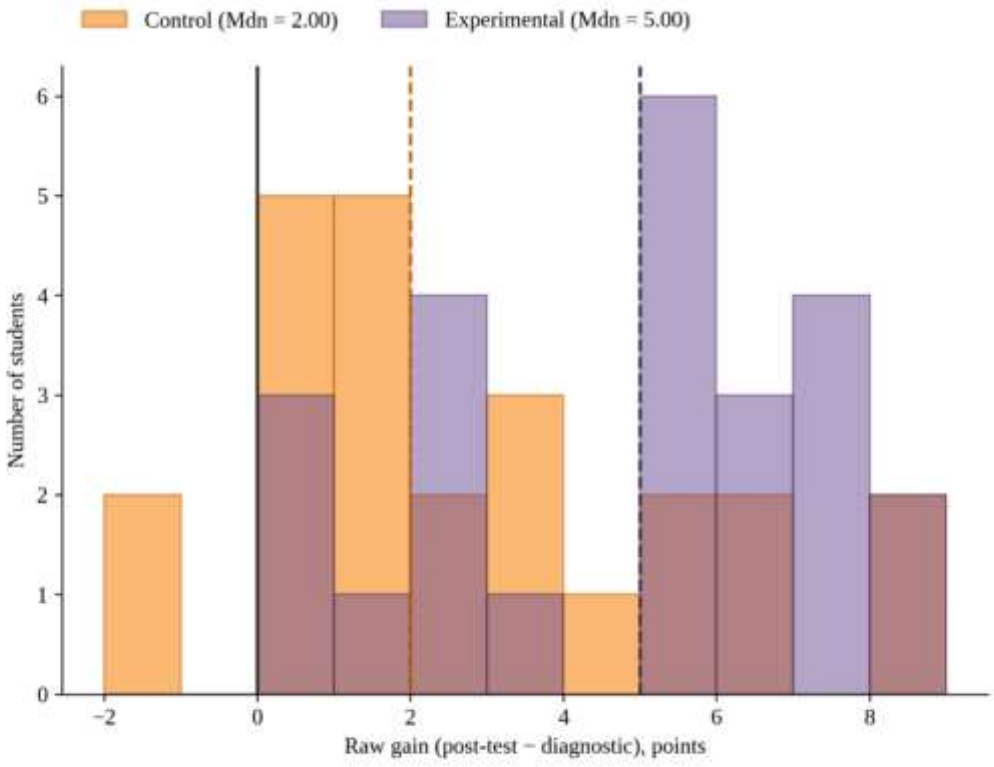
Table 4 revealed, however, that there was discrimination amongst the groups on the raw gain. The median gain between the two groups was 5.00 and 2.00 points, and this difference was of medium size, $U = 404.50$, $p = .016$ and $1.00 < \delta < 3.50$; the probability that a randomly selected student in the experimental group gained more than a randomly selected student in the control group was 0.702. It was therefore decided to support H_{12}. As shown (Figure 5), the control distribution was to the left of zero, as two students' scores dropped from 1.50 to zero and four students did not change, while the experimental distribution was exactly at zero and focused between 5.00 and 8.00 points.

Table 4. Between-group comparisons with the Mann–Whitney U test

Variable	U	z	p	Cliff's δ	95% CI	Rank-biserial r	V–D A	HL
Diagnostic	213.50	−1.544	.125	−0.259	[−0.578, 0.083]	−0.259	0.371	−1.375
Post-test	289.50	0.033	.982	0.005	[−0.313, 0.318]	0.005	0.503	0.000
Raw gain	404.50	2.411	.016	0.405	[0.089, 0.688]	0.405	0.702	2.500

Note. Own elaboration. The experimental group is entered first, so positive values favor it. z includes a tie correction. Confidence intervals for Cliff's δ were obtained by bootstrapping with 10,000 resamples. V–D A = Vargha–Delaney A statistic, the probability that an experimental observation exceeds a control observation. HL = Hodges–Lehmann median shift; for raw gain its 95% Moses interval is [1.00, 3.50].

Figure 5. Distribution of raw gains by group



Note. Own elaboration. Bars show the number of students in each one-point gain interval, where gain is the post-test grade minus the diagnostic grade. Dashed vertical lines mark each group's median and the solid dark line marks zero. Overlapping intervals are shown with transparency. The entire experimental histogram lies at non-negative values.

Overall, the combined (Kruskal–Wallis) test across the four conditions was highly significant, $\chi 2(3) = 28.731$, $p < .001$, $\varepsilon^2 = 0.280$. The likelihood of obtaining significant results (Table 5) for diagnostic to post-test contrasts did not become significant after adjustment, and was higher in the experimental group ($r = 0.649$ for the experimental group, vs. $r = 0.413$ for the control group), while the likelihood of obtaining significant results for the between group contrasts at matching moments did not achieve significance before adjustment. At the grand level of mean rankings, there was a slight, but insignificant, improvement of 22.61 places for the control group and 35.48 places for the experimental group. The difference between the conditions is in the trajectory, but not the end goal.

Table 5. Dunn multiple comparisons with Bonferroni adjustment across the four conditions

Comparison	R̄ 1	R̄ 2	z	p	p (Bonf.)	r
Control diag. – Control post.	38.83	61.44	−2.863	.004	.025	0.413
Control diag. – Exper. diag.	38.83	29.12	1.230	.219	1.000	0.177
Control diag. – Exper. post.	38.83	64.60	−3.264	.001	.007	0.471
Control post. – Exper. diag.	61.44	29.12	4.092	< .001	< .001	0.591
Control post. – Exper. post.	61.44	64.60	−0.401	.688	1.000	0.058
Exper. diag. – Exper. post.	29.12	64.60	−4.494	< .001	< .001	0.649

Note. Own elaboration. R̄ 1 and R̄ 2 are the mean ranks of the first and second condition, computed over all 96 observations. Standard errors include a tie correction. p (Bonf.) is multiplied by the number of comparisons (six). $r = |z|/\sqrt{N}$ for the observations involved.

Additionally, normalized gain was analyzed (Table 6), as this was also an advantage of the raw gain approach due to the mechanical advantage. The mean normalized of the classes was 0.731 in the experimental group, and 0.518 in the control group, showing a didactically substantial difference between them of 0.213. The individual normalized gain was not significantly different ($U = 233.50$; $p = .534$), perhaps because of the fact that its standard deviation in the control group (0.474) was almost double that of the experimental group (0.255) and the small sample size ($N = 41$) after excluding seven cases. The advantage of learning efficiency thus becomes apparent, describing the experiment, but not entailed from it.

Table 6. Normalized gain by group

Group	n	Class g	Individual g, M	SD	Mdn	Excluded cases
Control	20	0.518	0.578	0.474	0.750	4
Experimental	21	0.731	0.753	0.255	0.778	3

Note. Own elaboration. Individual normalized gain is g = (post-test − diagnostic)/(10 − diagnostic). Class g was computed from group means before excluding cases. Students with a diagnostic score of 10.00, for whom the denominator is zero, were excluded. The Mann–Whitney comparison of individual g yielded $U = 233.50$ ($p = .534$; rank-biserial $r = 0.112$).

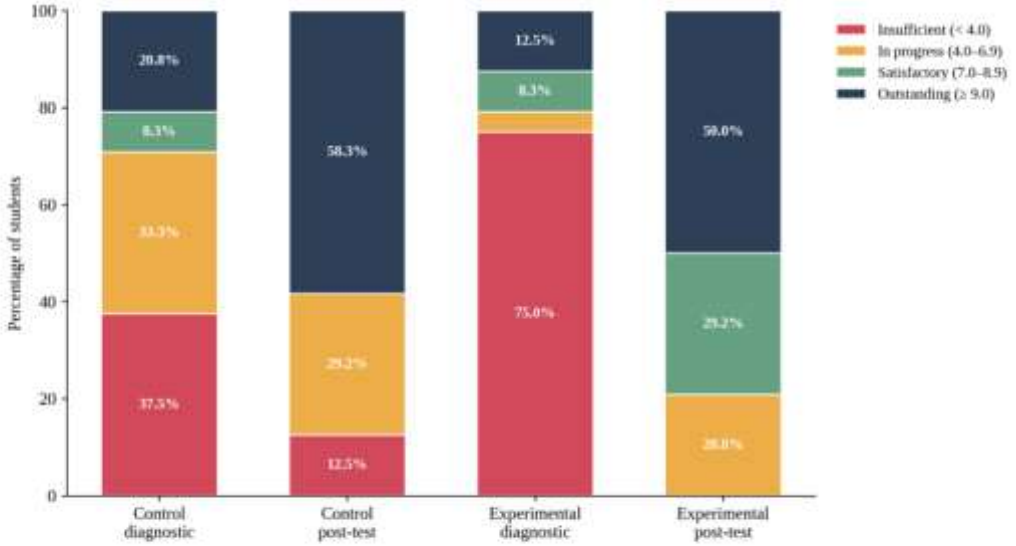
The most pedagogical reading is at the achievement level (Table 7, Figure 6). The three experimental students who failed the post-test were in the insufficient category, whereas no control student performed in this category; 79.2% of the experimental students, compared with 58.3% of the control, passed the exam. The odds across the four categories differed markedly, and this disparity was statistically highly significant with Fisher's exact test ($p = .015$): The control group had no satisfactory cases, displaying a bimodal distribution; the experimental group fell at the "high end" of the distribution on a continuous scale.

Table 7. Distribution of students by achievement level in each group and time point

Condition	Insufficient (< 4.0)	In progress (4.0–6.9)	Satisfactory (7.0–8.9)	Outstanding (≥ 9.0)	Passing (≥ 7.0)
Control – diagnostic	9 (37.5%)	8 (33.3%)	2 (8.3%)	5 (20.8%)	7 (29.2%)
Control – post-test	3 (12.5%)	7 (29.2%)	0 (0.0%)	14 (58.3%)	14 (58.3%)
Experimental – diagnostic	18 (75.0%)	1 (4.2%)	2 (8.3%)	3 (12.5%)	5 (20.8%)
Experimental – post-test	0 (0.0%)	5 (20.8%)	7 (29.2%)	12 (50.0%)	19 (79.2%)

Note. Own elaboration. Percentages are based on 24 students per condition. The last column combines Satisfactory and Outstanding. The test of independence between group and achievement level at post-test (four categories) yielded $\chi^2(3) = 10.487$ ($p = .015$). Fisher's exact test for the passing proportion yielded $OR = 2.714$ ($p = .212$).

Figure 6. Percentage composition of achievement levels by group and time point



Note. Own elaboration. Categories correspond to the grade ranges shown in the legend. Percentages are based on 24 students in each of the four conditions. Numeric labels for segments below 6% are omitted for legibility.

With H_{13}, there was unanimous consensus on the descriptive indicators (Table 8), with control variance nearly triple the experimental variance, the two other split ratios nearly doubling, and the bootstrap interval for the ratio of standard deviations excluding 1.00. However, they did not converge on inferential tests. The location difference is why the non-normal median-centered Levene and Fligner–Killeen tests failed to reject equal scales, whereas the mean-centered Levene and Ansari–Bradley tests did. The number of ties at 10.00 (absolute deviation from the control median is zero for 14 of 24 observations) does not affect the former tests but does affect the latter two. Therefore, we report descriptive support and robust support, as well as no conclusive inferential support, even though one might choose the test that confirmed the hypothesis, because we do not choose the test that confirms the hypothesis.

Table 8. Dispersion indicators for post-test grades and scale tests

Indicator or test	Control	Experimental	Ratio C/E	Inferential result
Variance	9.971	3.672	2.715	—
Standard deviation	3.158	1.916	1.648	95% CI [1.152, 2.399]
Coefficient of variation	0.416	0.231	1.802	—
Interquartile range	5.000	2.625	1.905	—
Total range	10.000	5.500	1.818	—
Observed minimum	0.000	4.500	—	—
Median-centered Levene	—	—	—	F = 1.482; p = .230
Mean-centered Levene	—	—	—	F = 13.329; p < .001
Fligner–Killeen	—	—	—	χ^2 = 0.042; p = .839
Ansari–Bradley	—	—	—	AB = 250.50; p = .003
Mann–Whitney on |x − Mdn|	—	—	—	U = 238.00; p = .297

Note. Own elaboration. Ratio C/E is the control value divided by the experimental value. The confidence interval for the ratio of standard deviations was obtained by bootstrapping with 10,000 resamples. The Ansari–Bradley test assumes equal medians, which does not hold here (10.00 vs. 8.75), so its result is confounded with the location difference and is reported only for completeness.

Results of robustness checks (Table 9) indicated that the gain advantage remained, and it grew even more robust when the seven students with the highest diagnostic scores were excluded (p = .010; rank

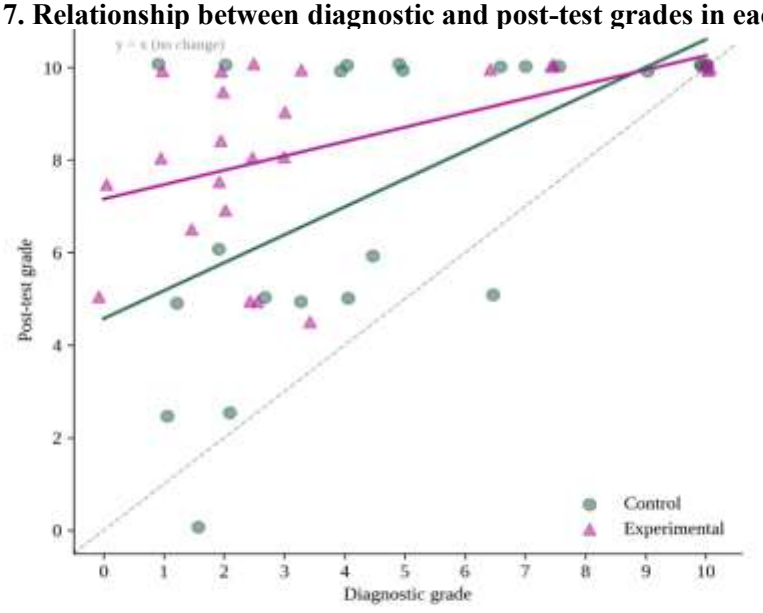
biserial r = 0.469). The absence of post-test differences also remained (p = .704). However, the mean residuals were in the predicted direction, and Quade's rank ANCOVA did not show a significant adjusted difference (t = 1.036; p = .306). There was a strong deviation from the unit-slope assumption inherent in using simple differences (the estimated rank slope = 0.508), which was not met. In the subgroup with diagnostic scores of 2.00 or lower, the post-test mean was almost 3 points higher in the experimental group (7.95) than in the control group (5.14); there were 10 and 7 cases, respectively, and equality was not significant (p = .140). In the subgroup with diagnostic scores of 7.00 or higher, all students in both groups scored the highest post-test value, as this part of the instrument provided no discrimination. The effect sizes, mean only the gain interval, exclude the null hypothesis, and its lower bound (0.089) falls below the threshold for a small effect size, which would indicate that there should not be a triumphalist reading of the numbers, from figure 8 or table 10.

Table 9. Robustness analyses of the main comparisons

Analysis	n control	n exper.	Statistic	p	Effect size
Gain, excluding diagnostic = 10	20	21	U = 308.50	.010	rank-biserial r = 0.469
Post-test, excluding diagnostic = 10	20	21	U = 224.50	.704	rank-biserial r = 0.069
Diagnostic, excluding diagnostic = 10	20	21	U = 147.50	.104	—
Rank ANCOVA (Quade)	24	24	t = 1.036	.306	slope = 0.508
Individual normalized gain	20	21	U = 233.50	.534	rank-biserial r = 0.112
Post-test, students with diagnostic ≤ 2	7	10	U = 50.50	.140	—
Gain, students with diagnostic ≤ 2	7	10	U = 52.00	.106	—

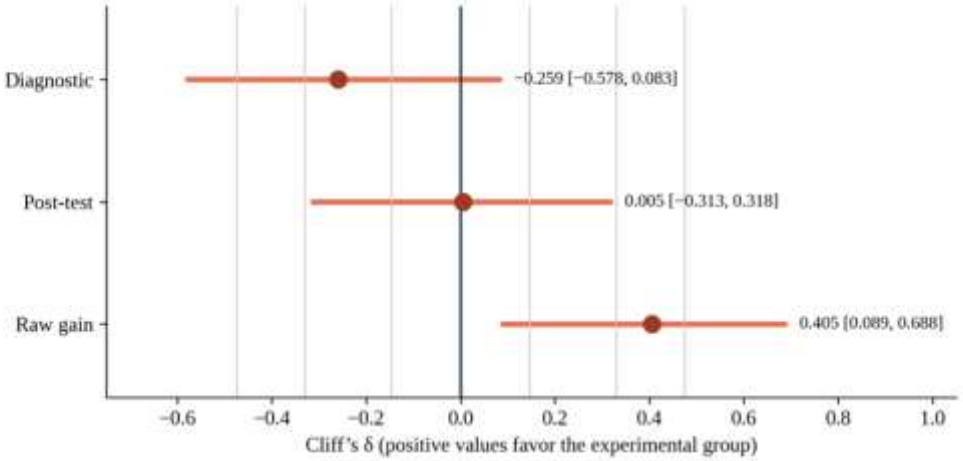
Note. Own elaboration. Quade's procedure ranks the covariate and the dependent variable over the pooled sample, fits a linear regression between the ranks, and compares residuals between groups; the reported slope belongs to that regression. The last two analyses are restricted to students with diagnostic scores of 2.00 or less, and their small size severely limits power.

Figure 7. Relationship between diagnostic and post-test grades in each group



Note. Own elaboration. Each marker represents one student; circles represent the control group and triangles the experimental group. Solid lines are each group's least-squares fits, included as a visual reference. The dashed diagonal represents no change. Minimal random jitter separates coincident observations. Spearman correlations between the two measurements were ρ = 0.612 (p = .002) in the control group and ρ = 0.485 (p = .016) in the experimental group.

Figure 8. Between-group effect sizes with 95% confidence intervals



Note. Own elaboration. Dots show point estimates of Cliff's δ and horizontal bars their 95% confidence intervals, obtained by bootstrapping with 10,000 resamples. Positive values favor the experimental group. Light gray vertical lines mark conventional thresholds (0.147 small; 0.330 medium; 0.474 large), and the dark line marks the null value.

Table 10. Summary of hypothesis tests

Hyp.	Statement	Decision	Main evidence
H01	The groups do not differ at the diagnostic	Retained	U = 213.50; p = .125; δ = −0.259
H11	The post-test median is higher in the experimental group	Rejected	U = 289.50; p = .982; δ = 0.005
H12	Raw gain is larger in the experimental group	Supported	U = 404.50; p = .016; δ = 0.405
H13	The applet reduces terminal dispersion	Descriptive, not inferential, support	Variance ratio = 2.715; Fligner p = .839

Note. Own elaboration. The decision on H11 means the alternative hypothesis found no empirical support, not that the control group was superior. For H13, descriptive indicators converge while the scale tests appropriate for non-normal distributions are not significant.

To reach these results, two apparently contradictory points must be kept in mind: the applet had a noticeable impact, and the applet had no impact on final performance. The conflict vanishes when we realize that the two statements point to different dependent variables. What is happening is described as a "ceiling effect." If over 50% of the distribution is above the scale maximum, the instrument will cease to measure. This deficit was not equal for both groups: the control group had a mean of four perfect diagnostic scores and seven students with scores higher than 7.00, which would be within reach of the ceiling no matter what they are told to do. Gain is partially free from this restriction by including each student's baseline, but it is mechanically weighted in favor of lower starting points, and the rank ANCOVA, by omitting the linear restriction (assuming just a rank vector difference from baseline), did not yield a significant difference. Evidence for the applet is thus directionally uniform but not uniformly strong across analytic models. The floor was raised for the experimental group (no student in the experimental group could have been below 4.50, no one regressed, and the size of terminal variance was better, being 2.72 times smaller). The best possible explanation of this pattern is the representativeness of the obstacle. If a student already has the ability to fluently convert between algebraic and graphical registers, then the student does not need to use the applet, because the student handles the conversion between the two registers in his or her head. The student who does not relies totally on outside sources and, without them, is incapable of reasoning from general principles to make decisions about any region other than the ones he or she has practiced. That is just the one in the lower tail of the control distribution (Duval, 2006; Trigueros & Martínez-Planell, 2010). While the design cannot demonstrate the mechanism, there is coherence between the two, and that serves as added support to the interpretation.

This research is closely linked to the literature in nuanced ways. It agrees with Takači et al. (2015) that calculus performance is better in a dynamic environment, but only because of the finding from their test, which they believe has a broader range, because of our Geogebra window, which of course cannot

be as broad as the one created by the authors of the article, or because of the dual nature of the double integral, which allows for more dichotomous testing. The measurements made by Radović et al. (2020) and Birgin and Topuz (2021) do not include delayed measurements, and there is no evidence here that the gain advantage will last; it is even possible that the students in the control group who have reached the measured ceiling with memorization will lose some of this gain more quickly. Chivai et al. (2022) reported similar post-test benefits but little persistence of 3D visualization problems. Our gain effect is well within the central point, as well as below Ji et al. (2024) or He et al. (2024), but if we report only post-test, it would be at the null point of the meta-analysis, and if we consider only gain, it would be at the medium point of the meta-analysis. Such a concentration of benefits among students with the lowest prior performance has not been observed at the level of meta-analysis, probably partly due to the difference in the metrics used and in part because of the effectiveness of the respective students. This kind of concentration of benefits among students with the lowest prior performance is the one that Wassie and Zergaw (2019) and Yohannes and Chen (2021) advocated for documenting – here's one for them to read.

The pattern demonstrated here, in which benefits are focused on those who scored lowest on prior performance, is precisely what Wassie and Zergaw (2019) and Yohannes and Chen (2021) called out for documenting – here's another one to read. Thirdly, the results are consistent with those of Ayala Chauvin and Luna Romero (2025), and we also extend their design by giving students the boost of learning where it was not offered as an automatic process: improved conversion between representations benefits students disproportionately, especially where it was not offered automatically. There are three implications regarding the teaching of multivariable calculus. In contrast, there is no evidence in our data to support a distinction in increased average achievement, but as the number of students receiving an insufficient grade drops to zero, a passing rate margin of 21 points between the two is strong evidence in a context of institutional problems of repetition. Second, the applet would not have had this effect alone: any instruction that emphasizes written predictions, verification, and reformulation, as done by Bedada and Machaba (2024), will have this effect, as noted by Takači et al. (2015), who suggest that technology has a role to play in teaching only if this happens within the instructional model teaching is based on. Third, though, an instrument where 58.3% in a group gets the highest score is not measuring the end result of the course towards what is in the target; more discriminatory instruments at the top, e.g., order-changing tasks requiring the decomposition of the region, modeling tasks from statements that are not so formalized, are still a future task.

The results are also relevant to quantitative bioscience education, but they were not directly obtained and can only be interpreted as hypotheses. Converting from a symbolic representation to a spatial or graphical representation of an object, as here, is the same problem that biology and biochemistry students face in enzyme kinetics. The Michaelis–Menten rate law, $v = V_{max}[S]/(K_m + [S])$, must be read two ways: as an equation and as a saturating curve whose shape is determined by two parameters; the integrated form is a rate vs. the time integral of the substrate concentration; metabolic flux is the accumulated amount of product from a pathway over time, and it is an integral of a rate. If the rate depends on two variables, such as the concentrations of a substrate and an inhibitor, then the rate is actually a surface, and the result over a range of conditions is exactly a double integral over a region of the concentration plane. Here, the task design is the same, but the prediction is made in writing, and after the student moves the sliders for Km, Vmax, or inhibitor concentration, he or she verifies this prediction, which will be explained. In our data, the greatest benefit was found at the lowest starting point, and since life-science students are often not as well-prepared in the mathematical sciences as are engineering students, a floor-raising effect may be more prominent in the life sciences.

4. Limitations and Scope for Future Research

The study has a few limitations, and the main constraint is the design. If differences between preexisting sections were not accounted for by randomization, they could be due to differences in group composition, session timing, or social dynamics between groups. This threat is minimized by checking for initial equivalence but not eliminated, and the prior difference may be of medium magnitude, as it falls within the confidence interval of the difference in interpretation (+/−0.578). The second limitation is a small or medium effect size, which would occur with 24 students per group; three of these subgroups (normalized gain, passing rate, and low-prior-performance subgroup) had non-significant results that might not reflect a lack of impact but rather a lack of power. For this reason, we report them without qualifications that may suggest a trend and consider them non-significant. The third limitation is the ceiling effect; the fourth is the absence of a delayed measurement to draw conclusions about retention; and the fifth is construct validity, because this instrument measures performance, not conceptual understanding, in solving double integrals, and a student might set up correct limits even if they do not

realize that the resulting integral represents a volume.
Finally, both groups had a similar class setup, with the same instructor and contact hours. However, for the experimental group, the setup was new and attention was heightened, so part of the effect may be attributed to novelty and increased attention. Four potential directions now seem to be priorities for future research. The first is the ability to reproduce on instruments that are not limited in their top note, thus restoring a level's discriminating ability. The second is an indirect measure taken at 8-12 weeks after training, that is, where learning across registers was mediated, to check whether this learning is more resistant to forgetting than memory for a procedure. The third is studies that are sufficiently powered for subgroup analyses, especially those focused on testing the hypothesis that the benefit lies among students who do not perform at an adequate level on the previous test or data point. The fourth approach is variation in the analytic model, such as quantile regression, which would capture the effects of treatment at various points of the distribution rather than at the distribution's center, particularly relevant to crossing distribution functions like those that occurred here. An active control would also help to separate out the effects of dynamic manipulation. Lastly, the implications drawn for quantitative bioscience education will need to be tested directly: A replication in a biochemistry or biology class using the same re-scaled Michaelis–Menten kinetics, re-scaled metabolic flux, and integrated rate laws as applets, and the same predict–check–explain tasks, would determine whether the pattern of having to rescale the learning content to raise the floor is alive and kicking in that population and content.

5. Conclusion

This study used non-parametric tests for independent groups to assess the effect of an applet on learning the double integral in Cartesian coordinates in two Civil Engineering sections. The groups did not differ significantly on the diagnostic, and both groups showed a significant improvement after six weeks of training. The standalone conclusion is that the applet had an impact on learning but not on the learning result. The experimental group had a significantly greater raw gain, a medium effect size, and a median gain score on a 10-point scale that was 2.5 points higher, a finding that remained true and somewhat stronger even after a sensitivity analysis of students who achieved maximum diagnostic scores. Final grades, on the other hand, not only did not differ but also showed one of the highest possible probabilities with a negligible effect size. The reason for this is the distributions' shapes. Both groups' CDFs intersect around 8.5 points, meaning that the experimental group has an advantage at the bottom of the graph and the control group has an advantage at the top, thus the two groups look equal at post-test, yet it is the result of the two advantages cancelling each other out. Therefore, the impact of the applet was an increase in the floor level and not the ceiling: none of the exposed students achieved scores lower than 4.50; one or more students achieved scores lower than would have been reached by any student in the control group; the floor height did not go down; the passing rate was raised by 21 percentage points; and the standard deviation (SD) for terminal variance was reduced from 1.36 to 1.03. Scale tests for non-normal data did not reach significance, and the hypothesis of reduced dispersion is asserted descriptively rather than statistically. The first methodological point of the study is that the dependent variable used in a technology-efficacy study determines the study's conclusion. The same 48 students yield null, medium, or non-significant effects when analyzed as final level, raw gain, or covariate-adjusted post-test, respectively. The three measures should therefore ALL be reported in primary studies, not just one. From a didactic perspective, a resource that does not go beyond the level but instead provides a means for avoiding failure at the level has significant didactic value in situations where iteration is the predominant issue, and this value is lost on measures of central position. Whether the same representational needs occur with other, similar concepts, such as learning about enzyme kinetics and metabolic flux among life science students, remains to be tested. It's not just a question of how much a class can improve; it's a question of who can improve in a class.

References

1. Ayala Chauvin, M. A., & Luna Romero, R. L. (2025). The use of GeoGebra for understanding implicit differentiation in engineering. Revista Científica Multidisciplinaria SAPIENTIAE, 8(16), 376–386. https://doi.org/10.56124/sapientiae.v8i16.021
2. Bedada, T. B., & Machaba, M. F. (2022). Investigation of student's perception learning calculus with GeoGebra and cycle model. Eurasia Journal of Mathematics, Science and Technology Education, 18(10), em2164. https://doi.org/10.29333/ejmste/12443
3. Bedada, T. B., & Machaba, M. F. (2024). The development of the cycle model and its effect on mathematics learning using GeoGebra mathematical software. Education Inquiry, 15(4), 503–526.

https://doi.org/10.1080/20004508.2022.2137260
4. Birgin, O., & Topuz, F. (2021). Effect of the GeoGebra software-supported collaborative learning environment on seventh grade students' geometry achievement, retention and attitudes. The Journal of Educational Research, 114(5), 474–494. https://doi.org/10.1080/00220671.2021.1983505
5. Brítez Silvero, J. de la C. (2016). Ethics in human research and the Ethics Committee (2). Revista Virtual de la Sociedad Paraguaya de Medicina Interna, 3(1), 8–10.
https://doi.org/10.18004/rvspmi/2312-3893/2016.03(01)08-010
6. Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. Journal of the American Statistical Association, 69(346), 364–367. https://doi.org/10.1080/01621459.1974.10482955
7. Carvalho, P. P. de, Silva, A. N. da, Silva, M. da C. V. da, & Rodrigues, W. F. da S. (2024). Exploring concepts of definite integrals in two variables using GeoGebra. Discover Education, 3(1), 1–27. https://doi.org/10.1007/s44217-024-00290-9
8. Castro-Molina, F. J. (2020). Ética vs. investigación [Ethics vs. research]. Cultura de los Cuidados, 24(57), 4–8. https://doi.org/10.14198/cuid.2020.57.01
9. Chan, K. K., & Leung, S. W. (2014). Dynamic geometry software improves mathematical achievement: Systematic review and meta-analysis. Journal of Educational Computing Research, 51(3), 311–325. https://doi.org/10.2190/EC.51.3.c
10. Chivai, C. H., Soares, A. A., & Catarino, P. (2022). Application of GeoGebra in the teaching of descriptive geometry: Sections of solids. Mathematics, 10(17), 3034.
https://doi.org/10.3390/math10173034
11. Clark-Wilson, A., Robutti, O., & Thomas, M. (2020). Teaching with digital technology. ZDM Mathematics Education, 52(7), 1223–1242. https://doi.org/10.1007/s11858-020-01196-0
12. Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics, 23(4), 351–361. https://doi.org/10.1080/00401706.1981.10487680
13. Dockendorff, M., & Solar, H. (2018). ICT integration in mathematics initial teacher training and its impact on visualization: The case of GeoGebra. International Journal of Mathematical Education in Science and Technology, 49(1), 66–84. https://doi.org/10.1080/0020739X.2017.1341060
14. Dunn, O. J. (1964). Multiple comparisons using rank sums. Technometrics, 6(3), 241–252. https://doi.org/10.1080/00401706.1964.10490181
15. Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. Educational Studies in Mathematics, 61(1–2), 103–131. https://doi.org/10.1007/s10649-006-0400-z
16. Gonscherowski, P., & Rott, B. (2022). How do pre-/in-service mathematics teachers reason for or against the use of digital technology in teaching? Mathematics, 10(13), 2345.
https://doi.org/10.3390/math10132345
17. Guerrero Chirinos, R. A., Torres Paqui, K. N., Silva Montoya, A. D., Cabrera Jiménez, A. E., Jiménez Rohoden, M. E., & Vacacela Ramón, D. O. (2025a). Contribution of innovative pedagogies to the development of algebraic thinking in middle general basic education students. Emergentes – Revista Científica, 5(2), 686–695. https://doi.org/10.60112/erc.v5.i2.442
18. Guerrero Chirinos, R. A., Delgado Fernández, J. R., Pacheco Molina, A., Vivanco Ureña, C. I., Reyes Carrión, J. P., & Vivanco Ureña, J. (2025b). Expert content validation: Interrater agreement and a standardized model for research instruments. Revista Latinoamericana de Educación, 3(3), e423. https://doi.org/10.53595/rle.v3.i3.015
19. He, A., Yuan, W., & Kiliçman, A. (2024). A meta-analysis on the effectiveness of dynamic mathematics software on K-12 students' mathematics learning. International Journal of Mathematical Education in Science and Technology. Advance online publication.
https://doi.org/10.1080/0020739X.2024.2375257
20. Ji, Z., Guo, K., & Song, S. (2024). Effects of dynamic mathematical software on students' performance: A three-level meta-analysis. Journal of Educational Computing Research, 62(4), 1035–1060. https://doi.org/10.1177/07356331241226594
21. Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621.
https://doi.org/10.1080/01621459.1952.10483441
22. Laderas Huillcahuari, E., Acori Flores, V., & Villa Pérez, L. (2023). Teaching differential and integral calculus assisted by GeoGebra software. Revista Latinoamericana de Investigación en Matemática Educativa, 26(3), 357–377. https://doi.org/10.12802/relime.23.2634
23. Martínez-Planell, R., & Trigueros, M. (2012). Students' understanding of the general notion of a function of two variables. Educational Studies in Mathematics, 81(3), 365–384.
https://doi.org/10.1007/s10649-012-9408-8
24. Pierce, R., & Stacey, K. (2010). Mapping pedagogical opportunities provided by mathematics

analysis software. International Journal of Computers for Mathematical Learning, 15(1), 1–20. https://doi.org/10.1007/s10758-010-9158-6
25. Quade, D. (1967). Rank analysis of covariance. Journal of the American Statistical Association, 62(320), 1187–1200. https://doi.org/10.1080/01621459.1967.10500925
26. Radović, S., Radojičić, M., Veljković, K., & Marić, M. (2020). Examining the effects of GeoGebra applets on mathematics learning using interactive mathematics textbook. Interactive Learning Environments, 28(1), 32–49. https://doi.org/10.1080/10494820.2018.1512001
27. Rasmussen, C., Marrongelle, K., & Borba, M. C. (2014). Research on calculus: What do we know and where do we need to go? ZDM Mathematics Education, 46(4), 507–515. https://doi.org/10.1007/s11858-014-0615-x
28. Sánchez Romero, J. E., Viramontes Campos, O., Viramontes Anaya, E., Guerrero Chirinos, R. A., Caizapanta Suárez, J. A., & Narváez Narváez, C. R. (2026). Artificial intelligence as a mediator of university learning: Promises, tensions, and horizons. Emergentes – Revista Científica, 6(2), 751–767. https://doi.org/10.60112/erc.v6.i2.791
29. Takači, Đ., Stankov, G., & Milanovic, I. (2015). Efficiency of learning environment using GeoGebra when calculus contents are learned in collaborative groups. Computers & Education, 82, 421–431. https://doi.org/10.1016/j.compedu.2014.12.002
30. Tatar, E., & Zengin, Y. (2016). Conceptual understanding of definite integral with GeoGebra. Computers in the Schools, 33(2), 120–132. https://doi.org/10.1080/07380569.2016.1177480
31. Thurm, D., & Barzel, B. (2020). Effects of a professional development program for teaching mathematics with technology on teachers' beliefs, self-efficacy and practices. ZDM Mathematics Education, 52(7), 1411–1422. https://doi.org/10.1007/s11858-020-01158-6
32. Thurm, D., & Barzel, B. (2022). Teaching mathematics with technology: A multidimensional analysis of teacher beliefs. Educational Studies in Mathematics, 109(1), 41–63. https://doi.org/10.1007/s10649-021-10072-x
33. Trigueros, M., & Martínez-Planell, R. (2010). Geometrical representations in the learning of two-variable functions. Educational Studies in Mathematics, 73(1), 3–19. https://doi.org/10.1007/s10649-009-9201-5
34. Wassie, Y. A., & Zergaw, G. A. (2019). Some of the potential affordances, challenges and limitations of using GeoGebra in mathematics education. Eurasia Journal of Mathematics, Science and Technology Education, 15(8), em1734. https://doi.org/10.29333/ejmste/108436
35. Weber, E., & Thompson, P. W. (2014). Students' images of two-variable functions and their graphs. Educational Studies in Mathematics, 87(1), 67–85. https://doi.org/10.1007/s10649-014-9548-0
36. Yohannes, A., & Chen, H.-L. (2021). GeoGebra in mathematics education: A systematic review of journal articles published from 2010 to 2020. Interactive Learning Environments. Advance online publication. https://doi.org/10.1080/10494820.2021.2016861
37. Zhang, Y., Wang, P., Jia, W., Zhang, A., & Chen, G. (2023). Dynamic visualization by GeoGebra for mathematics learning: A meta-analysis of 20 years of research. Journal of Research on Technology in Education. Advance online publication. https://doi.org/10.1080/15391523.2023.2250886
38. Ziatdinov, R., & Valles, J. R., Jr. (2022). Synthesis of modeling, visualization, and programming in GeoGebra as an effective approach for teaching and learning STEM topics. Mathematics, 10(3), 398. https://doi.org/10.3390/math10030398
39. Zulnaidi, H., Oktavika, E., & Hidayat, R. (2020). Effect of use of GeoGebra on achievement of high school mathematics students. Education and Information Technologies, 25(1), 51–72. https://doi.org/10.1007/s10639-019-09899-y