Revista Multidisciplinaria Perspectivas Investigativas
Multidisciplinary Journal Investigative Perspectives
Vol. 6(4), 35-43, 2026
https://doi.org/10.62574/rmpi.v6i4.526

IOBAS strategy with artificial intelligence chatbots and mathematics learning in lower secondary education

Estrategia IOBAS con chatbots de inteligencia artificial y aprendizaje de las matemáticas en básica secundaria

Johann Verney Méndez-Gamba
johann-v.mendez-g@up.ac.pa
Universidad de Panamá, Panamá, Provincia de Panamá, Panamá
https://orcid.org/0009-0008-6332-4850

ABSTRACT

The study analyzes the incidence of artificial intelligence chatbots integrated into the IOBAS teaching strategy on mathematics learning among lower secondary students of a technical school in Samacá, Colombia. A sequential explanatory mixed design was used, with a quasi-experimental pretest-posttest component (30 students in the experimental group and 34 in the control group), a problem-solving rubric, a perception survey, classroom observations and learning artifacts. The overall gain was 8.60 points in the experimental group and 5.46 in the control group (g = 1.21), and the difference remained after adjusting for the pretest (η^2p = .251), with greater consistency in the conceptual and interpretative dimensions than in the procedural one. It is concluded that the integrated condition was associated with better performance, although the effect cannot be attributed to the chatbots in isolation.

Descriptors: artificial intelligence; mathematics; teaching methods. (Source: UNESCO Thesaurus).

RESUMEN

El estudio analiza la incidencia de chatbots de inteligencia artificial integrados a la estrategia didáctica IOBAS en el aprendizaje de las matemáticas de estudiantes de básica secundaria de una institución técnica de Samacá, Colombia. Se empleó un diseño mixto secuencial explicativo, con un componente cuasiexperimental pretest-postest (30 estudiantes en el grupo experimental y 34 en el control), una rúbrica de resolución de problemas, una encuesta de percepción, observaciones de aula y artefactos de aprendizaje. La ganancia global fue de 8,60 puntos en el grupo experimental y de 5,46 en el control (g = 1,21), y la diferencia se mantuvo tras ajustar por el pretest (η^2p = ,251), con mayor consistencia en las dimensiones conceptual e interpretativa que en la procedimental. Se concluye que la condición integrada se asoció con un mejor desempeño, sin que el efecto pueda atribuirse a los chatbots de forma aislada.

Descriptores: inteligencia artificial; matemáticas; métodos de enseñanza. (Fuente: Tesauro UNESCO).

Received: 5/10/2026. Reviewed: 12/10/2026. Accepted: 18/10/2026. Published: 5/11/2026.
Research Articles Section

INTRODUCTION

Learning mathematics in lower secondary education depends to a large extent on students' ability to solve problems, that is, to understand a situation, recognize the relevant information, choose a strategy and justify the result. Along these lines, Benítez-Pérez, et al., (2023) link comprehension of the problem statement with the inductive and deductive processes involved in solving it, while Pérez-González, et al., (2021) propose a teaching strategy aimed at directing the teaching and learning process of mathematical problem solving. From this perspective, the difficulties are not limited to calculation but extend to the initial phases of reading, organization and planning, which justifies the interest in strategies that make each moment of the process visible.

One such strategy is IOBAS, a problem-solving proposal of Peruvian origin that structures the work into five moments: identification of the problem, organization of knowledge, search for and choice of strategies, application, and synthesis or socialization of results. Its grounding in problem-based learning makes it compatible with tools that offer conversational support, so that artificial intelligence chatbots can act as scaffolding in each phase. With regard to scaffolding, van de Pol, et al., (2019) analyzed how students take up teacher support in small-group work, and Sweller, et al., (2019) argue that human cognitive architecture requires structuring the information presented to those learning new content.

At the same time, research on generative artificial intelligence in mathematics education is growing rapidly. Debets, et al., (2025) systematically reviewed the objectives, technological and theoretical basis, evaluation criteria and impacts of chatbots in education; Pepin, et al., (2025) mapped studies on ChatGPT in mathematics teaching; and Rutherford, et al., (2025) explored teachers' use of this tool to design lessons with motivational support. For their part, the meta-analysis by Wang & Fan (2025) reports favorable effects of ChatGPT on performance, perception and higher-order thinking, with differences depending on the characteristics of the implementation. However, automated feedback does not in itself guarantee better learning: Kluger & DeNisi (1996) showed that its effects are variable and that in some cases they lower performance, and Wisniewski, et al., (2020) found that it is more effective when it provides information about the task and the process, in line with Shute's (2008) review of formative feedback.

In this context, evidence that combines a structured problem-solving strategy with chatbots, compares groups and at the same time examines classroom processes is scarce in the sources reviewed, and studies tend to describe potential rather than compare conditions. Based on the above, this article, derived from a completed doctoral thesis, aims to analyze the incidence of artificial intelligence chatbots, integrated into the IOBAS teaching strategy, on mathematics learning among lower secondary students at the Institución Educativa Técnica Nacionalizada de Samacá, Colombia.

METHOD

The research adopted the pragmatic paradigm and a sequential explanatory mixed design (QUAN → qual), in which the quantitative phase preceded and guided the qualitative one (Creswell & Plano Clark, 2018). The quantitative component was quasi-experimental, with pretest and posttest measurement in an experimental group, which worked with IOBAS mediated by chatbots, and a control group, which received traditional teaching equivalent in time and objectives. The groups were intact classes and were not randomly assigned. The qualitative component examined classroom processes through observations and learning artifacts, and integration was carried out with a joint display matrix and meta-inferences.

The study was conducted at the Institución Educativa Técnica Nacionalizada de Samacá, Colombia, with lower secondary students. The performance test included 30 students from the experimental group with complete pretest and posttest records, and 34 from the control group in the posttest, 33 of whom also had a pretest. One control record without pretest scores was excluded only from the analyses that required the baseline measurement. Missing data were not imputed. Regarding ethical aspects, consent was obtained from parents or guardians and assent from the students, voluntary participation was ensured, and the data and traces of interaction with the chatbots were pseudonymized.

In the experimental group, students solved problem situations following the five phases of IOBAS and consulted chatbots as support, under rules of use that prioritized teacher verification of the responses and the gradual withdrawal of support. The control group worked on the same content with the usual methodology. The activities focused mainly on the linear function and the straight line in the Cartesian plane, with attention to slope, intercept, and graphical and algebraic representation.

The quantitative instruments were a performance test administered before and after the intervention, whose overall score is the sum of the conceptual, procedural and interpretative dimensions; an analytical IOBAS problem-solving rubric with five criteria (identification of the problem, organization of knowledge, procedure and application of strategies, verification, and justification) scored from 1 to 4, so that the overall score ranges from 5 to 20 and is classified into very low (5 to 8), low (9 to 12), medium (13 to 16) and high (17 to 20) levels; and an 11-item perception survey on a 1-to-5 scale, administered to the experimental group. The qualitative sources were an observation guide for classroom discourse and practice, which considered the ICAP framework, sociomathematical norms, scaffolding and feedback, errors, and the interaction among student, chatbot and teacher, together with the analysis of learning artifacts and their interaction traces.

Data were described using mean, standard deviation and median. Between-group comparisons were performed with Welch's t-test and, complementarily, with the Mann-Whitney U test, after checking normality with Shapiro-Wilk. Within-group changes were tested with the paired-samples t-test and with Wilcoxon, and Cohen's d for repeated measures (dz) and Hedges' g for the difference in gains were estimated. To control for the initial level, an analysis of covariance (ANCOVA) was fitted with the posttest as the dependent variable, group as the factor and the pretest as the covariate, after checking the homogeneity of slopes, and partial eta squared (η^2p) was reported. The internal consistency of the survey was estimated with Cronbach's alpha, and its relationship with individual gains with Spearman's ρ coefficient. The significance level was $\alpha = .05$. The qualitative analysis compared the evidence from the observations and artifacts on the basis of the instrument categories, without introducing categories foreign to the records.

RESULTS

The results are presented following the analytical sequence: initial performance, final performance and gains, the pretest-adjusted analysis, the IOBAS rubric, student perceptions and qualitative findings.

Initial and final performance

In the pretest, both groups started from a low level. The experimental group obtained an overall mean of 1.07 (SD = 0.89) and the control group 1.62 (SD = 1.50). The difference did not reach significance with Welch's t-test, $t(52.90) = -1.81$; $p = .075$, or with Mann-Whitney, $U = 393.0$; $p = .162$. No significant differences were found in the conceptual (0.84 vs. 0.90) or procedural (0.10 vs. 0.39) dimensions either. In the interpretative dimension, by contrast, Welch's test showed a difference between the means (0.13 vs. 0.33), $t(54.48) = -2.13$; $p = .038$, which Mann-Whitney did not confirm, $U = 376.0$; $p = .072$. Descriptively, the control group had higher means in the overall score and in all three dimensions.

In the posttest, the experimental group reached an overall mean of 9.67 (SD = 2.49) and the control group 6.94 (SD = 2.94). The difference was statistically significant, $t(61.90) = 4.02$; $p < .001$, and was confirmed with Mann-Whitney, $U = 782.0$; $p < .001$. By dimension, the conceptual score was 3.36 (SD = 0.87) vs. 2.01 (SD = 0.92), $t(61.70) = 6.07$; $p < .001$, and the interpretative score 2.88 (SD = 1.14) vs. 2.05 (SD = 1.29), $t(62.00) = 2.71$; $p = .009$. In the procedural dimension, the means were 3.44 (SD = 1.26) and 2.89 (SD = 1.33), a difference that did not reach significance, $t(61.67) = 1.70$; $p = .095$; $U = 651.0$; $p = .059$.

Pretest-posttest changes and gains

Both groups improved between the two measurements. In the experimental group, the overall score increased by 8.60 points (SD = 2.36; 95% CI [7.72, 9.49]), $t(29) = 19.93$; $p < .001$; $dz = 3.64$. In the control group, the increase was 5.46 points (SD = 2.75; 95% CI [4.48, 6.43]), $t(32) = 11.41$; $p < .001$; $dz = 1.99$; this group recorded some negative individual changes, with a minimum of −0.88. The Wilcoxon tests agreed with these decisions. The values are summarized in Table

1.

Table 1. *Overall performance test score in the pretest, the posttest and gain by group*.

Group	n pre	Pretest M (SD)	n post	Posttest M (SD)	n Δ	Gain Δ M (SD)
Experimental	30	1.07 (0.89)	30	9.67 (2.49)	30	8.60 (2.36)
Control	33	1.62 (1.50)	34	6.94 (2.94)	33	5.46 (2.75)

Note: developed in-house. M = mean; SD = standard deviation; Δ = posttest − pretest. The control group gain was calculated with the 33 cases that had both measurements, whose posttest mean was 7.08 (SD = 2.87).

The direct comparison of gains showed a difference of 3.15 points in favor of the experimental group (95% CI [1.86, 4.44]), t(60.83) = 4.88; p < .001; g = 1.21, a result supported by Mann-Whitney, U = 801.5; p < .001. As shown in Table 2, the difference was significant in all three dimensions, with the largest magnitude in the conceptual dimension and the smallest in the procedural one.

Table 2. *Comparison of pretest-posttest gains between groups, by dimension*.

Variable	Experimental Δ M (SD)	Control Δ M (SD)	Difference	95% CI	t	p	g
Overall score	8.60 (2.36)	5.46 (2.75)	3.15	[1.86, 4.44]	4.88	< .001	1.21
Conceptual	2.52 (0.97)	1.14 (0.99)	1.38	[0.88, 1.87]	5.58	< .001	1.39
Procedural	3.34 (1.19)	2.54 (1.44)	0.81	[0.14, 1.47]	2.42	.019	0.60
Interpretative	2.75 (1.13)	1.78 (1.22)	0.96	[0.37, 1.56]	3.24	.002	0.81

Note: developed in-house. Δ = posttest − pretest; M = mean; SD = standard deviation; 95% CI = confidence interval of the difference; t = Welch's t (df between 60.5 and 61.0); g = Hedges' g. n = 30 in the experimental group and 33 in the control group.

Analysis adjusted for baseline performance

Regression slopes were homogeneous for the overall score, F(1, 59) = 0.13; p = .715, and for all three dimensions, so the models were estimated without the interaction term. For the overall score, the difference persisted after controlling for the pretest, F(1, 60) = 20.14; p < .001; η^2p = .251, with adjusted means of 9.88 for the experimental group and 6.90 for the control group (difference of 2.98; 95% CI [1.65, 4.31]). The model explained 27.8% of the variability in the posttest. Table 3 presents the results by dimension.

Table 3. *ANCOVA results on the posttest, with the pretest as covariate*.

Variable	Adjusted M experimental	Adjusted M control	Difference	95% CI	F(1, 60)	p	η^2p
Overall score	9.88	6.90	2.98	[1.65, 4.31]	20.14	< .001	.251
Conceptual	3.37	2.03	1.33	[0.89, 1.77]	36.42	< .001	.378
Procedural	3.50	2.87	0.63	[−0.03, 1.29]	3.65	.061	.057
Interpretative	2.95	2.04	0.91	[0.29, 1.53]	8.61	.005	.125

Note: developed in-house. M = mean; 95% CI = confidence interval of the adjusted difference; η^2p = partial eta squared. n = 30 in the experimental group and 33 in the control group.

In the procedural dimension, the group effect did not reach conventional significance (p = .061), the confidence interval included zero and the residuals departed from normality (W = .950; p = .012), so this result differs from the one obtained when comparing gains (p = .019). In the overall model and in those for the conceptual and interpretative dimensions, the residuals showed no departures from normality or heterogeneity of variances.

Performance on the IOBAS rubric

The rubric gathered 26 records of classroom activities, which are not equivalent to 26 independent students, since some correspond to individual students and others to pairs or groups. Their mean overall score was 12.96 (SD = 3.97; median = 12.50), with values between 6 and 20. Four records (15.4%) were at the very low level, nine (34.6%) at the low level, six (23.1%) at the medium level and seven (26.9%) at the high level. The criteria with the lowest means were identification of the problem and organization of knowledge, while verification and justification had the highest (Table 4).

The records came from two administration dates, 15 from May 20, 2026 and 11 from May 21,

2026. The mean overall score was 10.20 (SD = 2.21) on the first date and 16.73 (SD = 2.37) on the second, and the difference appeared in all five criteria. Since the units observed were not equivalent across dates, this comparison is only descriptive and does not constitute a paired measurement.

Table 4. *IOBAS rubric scores by criterion and administration date*.

Criterion	Total M (SD)	Median	May 20 M	May 21 M
Identification of the problem	2.12 (1.11)	2.00	1.40	3.09
Organization of knowledge	2.31 (1.12)	2.00	1.60	3.27
Procedure and strategies	2.62 (0.75)	3.00	2.20	3.18
Verification	2.92 (0.80)	3.00	2.47	3.55
Justification	3.00 (1.10)	3.00	2.53	3.64
Overall score (5 to 20)	12.96 (3.97)	12.50	10.20	16.73

Note: developed in-house. M = mean; SD = standard deviation. Each criterion was scored from 1 to 4. The total corresponds to 26 records, 15 from May 20 and 11 from May 21, 2026, which are not equivalent to each other.

Student perceptions

The survey gathered 25 complete forms, with no missing values and an internal consistency of α = .819. Codes 16 and 33 appeared twice and one of the codes had no match in the experimental group database, so the forms were not assumed to correspond to 25 different students. The pooled mean of the 275 responses was 4.06; 25.1% corresponded to category 5, 57.5% to category 4, 16.0% to category 3 and 1.5% to category 2, with no responses in category 1, so that 82.5% were concentrated in categories 4 and 5. Item means ranged from 3.68 to 4.32 and the median was 4 for all items (Table 5).

Table 5. *Descriptive results of the perception survey on IOBAS and the use of chatbots*.

Item	Abbreviated content	M	SD
1	Understanding of mathematical topics through chatbots	4.16	0.75
2	Organization of prior knowledge through IOBAS	3.88	0.78
3	Choice of strategies through chatbots	4.12	0.60
4	Dynamism of learning with chatbots	4.12	0.67
5	Self-assurance in applying problem-solving strategies	3.96	0.68
6	Learning through the socialization of results	3.68	0.63
7	Usefulness and personalization of chatbot feedback	4.04	0.61
8	Positive attitude toward learning mathematics	4.00	0.65
9	Confidence in using technological tools	4.32	0.69
10	Perceived improvement in learning through IOBAS	4.16	0.75
11	Perceived improvement in learning through AI and chatbots	4.24	0.60

Note: developed in-house. M = mean; SD = standard deviation. Scale from 1 to 5; 25 complete forms.

Confidence in using technological tools (item 9) obtained the highest mean and socialization of results (item 6) the lowest. When linking the responses to individual results, an unambiguous match was achieved for 22 participants; for the two duplicated codes, responses were averaged by item, and the alternative use of the first or the second response did not change the significance. The overall perception score had a mean of 4.08 (SD = 0.41). Its correlation with total gain was low and non-significant ($\rho = .169$; $p = .451$), as were its correlations with the conceptual ($\rho = -.219$; $p = .327$), procedural ($\rho = .271$; $p = .222$) and interpretative ($\rho = .328$; $p = .137$) gains.

Qualitative findings

The observation guide contains the records of sessions 01, 02, 03 and 05; the record for session 04 was not identified. In addition, the teacher and the observer appear under the same name, although the instrument called for non-participant observation. In the first session, conducted without a chatbot, passive participation predominated, focused on listening and taking notes. The second session recorded an exchange of ideas about strategies, use of the chatbots as support and a predominantly constructive level, with justifications that were still partial and correct proposals alongside incorrect ones. Session 03 maintained the constructive level with more peer discussion and, in session 05, the predominant level was interactive, with justification of answers, evaluation of strategies and reasoned validation of results.

Appropriation of the IOBAS phases was gradual. In the first session, all five phases were rated at

level 1, given its introductory nature. In session 02, identification and search were rated at level 1 and organization and synthesis at level 2, with uncertainty about what had to be done in each phase. In session 03, all five phases reached level 2, and session 05 recorded level 3 in identification, search, application and synthesis, and level 4 in organization. The artifacts showed variability: some products from May 20 showed low levels of identification and organization and procedures without a documented sequence, while others from May 21 reached levels 3 and 4, with complete organization of the data, recognition of the slope and the intercept, and consistent argumentation of the solutions.

The use of the chatbots was heterogeneous. In an initial approach, several students entered the problem statements to obtain the solution and, in one artifact, that query was classified as having low relevance, partial use of the response, medium dependence and low reworking. Other groups asked whether the procedure for obtaining the equation of a line was correct, asked about the meaning of the slope, about vertical, horizontal, parallel and perpendicular lines, or about whether their tables of values were correct, with high relevance, low dependence and high reworking in the latter case. In session 05, the questions were more specific and explanatory and queries to obtain direct solutions decreased, although some students continued to turn to the tool without first working through the process.

Specific difficulties persisted. Problems were observed in associating the slope with the rate of change, distinguishing the initial condition and relating it to the intercept, handling signs and units, contextualizing results and translating between the graphical representation and the equation. In session 05, conceptual, procedural and interpretative errors were classified as low frequency, while representational errors remained at medium frequency, mainly due to the scales of the axes. As for mediation, the teacher's direct explanations were frequent in the first session and of low frequency in the fifth, when guiding questions and formative feedback became more prominent.

DISCUSSION

The results indicate that the condition that integrated IOBAS and chatbots was associated with higher mathematical performance than traditional teaching. Both groups improved, but the experimental group did so by 3.15 points more ($g = 1.21$), and the difference remained after controlling for the pretest ($\eta^2p = .251$), so it does not depend solely on the initial advantage or disadvantage. This direction coincides with that of the meta-analysis by Wang & Fan (2025), which reports favorable effects of ChatGPT on performance, and with the review by Debets, et al., (2025), which places chatbots as a field with documented impacts. However, the experimental condition combined the IOBAS strategy, peer work, teacher mediation and chatbots, and the quasi-experimental design does not allow the contribution of each component to be separated, which is why the difference should be read as an association with the complete pedagogical configuration and not as an effect attributable to artificial intelligence. In addition, the groups were not necessarily equivalent at the outset, since the control group showed higher means and a discrepancy between tests in the interpretative dimension, and the absence of significance in the pretest is not equivalent to a test of equivalence.

Progress was not uniform across dimensions. The evidence was more consistent in the conceptual ($\eta^2p = .378$) and interpretative ($\eta^2p = .125$) dimensions, whereas the procedural dimension depended on the analysis: the gain was greater in the experimental group ($p = .019$), but the ANCOVA did not reach significance ($p = .061$) and the residuals departed from normality. Therefore, a conclusive procedural effect cannot be claimed. The qualitative records are consistent with this caution, since operational errors decreased but mistakes in signs, units and sequencing of steps continued. A plausible reading, not verified in this design, is that conversational support favors the explanation and interpretation of relationships more than the autonomous execution of procedures, an aspect that Sweller, et al., (2019) allow us to relate to the need to withdraw support gradually as expertise increases.

The differences between the IOBAS phases provide information that the final test result does not offer. On the rubric, identification of the problem (2.12) and organization of knowledge (2.31) obtained the lowest means, and the artifacts showed students performing appropriate operations without making explicit either the objective or the organization of the data. This pattern is consistent with the importance that Benítez-Pérez, et al., (2023) attach to comprehension of the problem statement and with the proposal by Pérez-González, et al., (2021) to direct problem

solving as a process, so that a correct result is not equivalent to an explicit process. The improvement between May 20 and May 21 (from 10.20 to 16.73) is compatible with a progressive appropriation of the strategy, although the units observed were not equivalent and it cannot be interpreted as individual change.

Regarding the chatbots, the records show different functions: obtaining the solution, guiding the procedure, clarifying concepts and verifying one's own work. This heterogeneity offers a plausible explanation, without causal demonstration, of why the favorable perception (4.06 out of 5) was not associated with individual gain ($\rho = .169$; $p = .451$), since two students may rate the tool equally and use it in very different ways. This is consistent with the literature on feedback, according to which its effects vary and sometimes impair performance (Kluger & DeNisi, 1996), and its effectiveness increases when it provides guidance on the task and the process (Wisniewski, et al., 2020). Along the same lines, Pepin, et al., (2025) and Rutherford, et al., (2025) show that the value of ChatGPT in mathematics education depends on how it is incorporated into teaching, and van de Pol, et al., (2019) highlight the role of teacher support in group work, which is consistent with the observed shift from direct explanations toward guiding questions. Moreover, the survey data describe ratings rather than an objective measure of learning, so the 82.5% of responses in categories 4 and 5 cannot be used as evidence of improvement.

The transition from passive participation to constructive and interactive levels can be read through the ICAP framework, according to which the interactive and constructive modes are associated with better outcomes than the passive and active modes (Chi & Wylie, 2014). However, the observations describe classroom dynamics rather than individual scores, session 04 was not recorded and the observer was linked to the implementation, so this correspondence is interpretative. The limitations include the non-random formation of the groups, baseline differences, losses and duplicates in some databases, the mixed recording unit of the rubric, the small size of the linked sample for the correlations and the fact that a single institution was involved, so the results should not be generalized to other grades, content or contexts without additional evidence.

CONCLUSIONS

The integration of artificial intelligence chatbots into the IOBAS strategy, accompanied by teacher mediation and peer work, was associated with a greater increase in the mathematical performance of lower secondary students in Samacá compared with traditional teaching: the overall gain was 8.60 points in the experimental group and 5.46 in the control group, and the difference persisted after adjusting for the pretest ($\eta^2p = .251$), with greater consistency in the conceptual and interpretative dimensions than in the procedural one. The rubric and the artifacts showed that the greatest difficulties were concentrated in identification of the problem and organization of knowledge, and the qualitative records revealed increasing participation and uses of the chatbots that ranged from requesting answers to clarification and verification. Student ratings were favorable but were not significantly related to individual gains. These results correspond to a composite pedagogical configuration, whose components could not be separated, and to a quasi-experimental design with baseline differences, so they support a situated assessment of the intervention rather than the general superiority of chatbots. The contribution of the study lies in showing that the value of chatbots depends on how they are integrated into the problem-solving process, subject to the student's explanation, verification and justification.

FINANCING

Non-monetary.

CONFLICT OF INTEREST

There is no conflict of interest with any individuals or institutions related to this research.

ACKNOWLEDGMENTS

To the students and teachers of the Institución Educativa Técnica Nacionalizada de Samacá, for their participation, and to Dr. César Ignacio Báez Quintero, supervisor of the doctoral thesis from which this article derives.

REFERENCES

Benítez-Pérez, A. A., García-Rodríguez, M. L., & Flores-Cervantes, C. (2023). An approach to text comprehension in mathematical problem solving considering inductive and deductive processes: The case of Bobby. *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 13*(26). https://doi.org/10.23913/ride.v13i26.1423

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Debets, T., Banihashem, S. K., Joosten-Ten Brinke, D., Vos, T. E. J., Maillette de Buy Wenniger, G., & Camp, G. (2025). Chatbots in education: A systematic review of objectives, underlying technology and theory, evaluation criteria, and impacts. *Computers & Education, 234*, 105323. https://doi.org/10.1016/j.compedu.2025.105323

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Pepin, B., Buchholtz, N., & Salinas-Hernández, U. (2025). A scoping survey of ChatGPT in mathematics education. *Digital Experiences in Mathematics Education, 11*, 9–41. https://doi.org/10.1007/s40751-025-00172-1

Pérez-González, A., Quero-Méndez, O. N., & Bravo-Viera, J. L. (2021). Didactic strategy for teaching how to direct the teaching-learning process of mathematical problem solving. *Revista Educación, 45*(1). https://doi.org/10.15517/revedu.v45i1.42112

Rutherford, T., Rodrigues, A., Duque-Baird, S., Veng, S., Mykyta-Chomsky, R., Cao, Y., Chisholm, K., & Bergwall, E. (2025). "I just think it is the way of the future": Teachers' use of ChatGPT to develop motivationally-supportive math lessons. *Computers & Education: Artificial Intelligence, 8*(3), 100367. https://doi.org/10.1016/j.caeai.2025.100367

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

van de Pol, J., Mercer, N., & Volman, M. (2019). Scaffolding student understanding in small-group work: Students' uptake of teacher support in subsequent small-group interaction. *Journal of the Learning Sciences, 28*(2), 206–239. https://doi.org/10.1080/10508406.2018.1522258

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications, 12*, 621. https://doi.org/10.1057/s41599-025-04787-y

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

Copyright: 2026 By the authors. This article is open access and distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license https://creativecommons.org/licenses/by-nc-sa/4.0/