SAMYRAD 2026

SMART ALGORITHMS & MODELS: YOUR REACH TO ADVANCED DIALOGUES · 5–6 OCTOBER 2026 · SEVILLE, SPAIN

SAMYRAD OPEN-ACCESS PROCEEDINGS | PUBLISHER: INTELLIGENT NUDGE

PAPER • OPEN ACCESS

From Automation to Judgment: Evaluating an AI-Augmented Monte Carlo Instructional Architecture for Responsible Managerial Decision-Making

Roberto Gómez Tobías, Giovanni María Baldini

(email. roberto.gt@tec.mx) and (email. gbaldini@tec.mx)

International Business Department, Tecnológico de Monterrey Monterrey, Mexico

Article DOI: 10.68286/samyrad2026.363

ABSTRACT

Generative artificial intelligence can help business students produce analyses, but learning also requires them to examine uncertainty and justify decisions. This study evaluates whether adding structured risk modeling and reflection to a common AI-supported managerial task is associated with stronger responsible judgment. Responsible Augmented Business Judgment (RABJ) means using AI for analytical support while retaining accountability for assumptions, uncertainty, stakeholder consequences, and ethically defensible decisions. A mixed-methods quasi-experimental study involved 120 undergraduates in 24 teams across four sections. Both conditions shared the case, roles, faculty guidance, and executive recommendation. The full intervention additionally used Monte Carlo simulation and a matrix for reflecting on present and future rational and emotional considerations. Evidence included a 16-item pre-post survey, team level performance rubrics, expert content-validation evidence, and qualitative coding of anonymized reflection summaries. Baseline-adjusted ANCOVA favored the full intervention on the overall composite, $B = 0.557$, 95% CI [0.482, 0.632], $p < .001$; the strongest adjusted difference occurred in metacognition and emotional foresight, $B = 0.793$. Team performance showed the same directional pattern (3.47 vs. 2.93), with high interrater reliability, $ICC(2,k) = .954$. Findings suggest that structured uncertainty analysis and reflection can support accountable use of AI in business education.

KEYWORDS Artificial intelligence, Monte Carlo simulation, responsible managerial judgment, business education, instructional design, ethical decision-making, uncertainty

How to cite: Gómez Tobías, R., & Baldini, G. M. (2026). From automation to judgment: Evaluating an AI-augmented Monte Carlo instructional architecture for responsible managerial decision-making. SAMYRAD 2026 Open-Access Proceedings. Intelligent Nudge. https://doi.org/10.68286/samyrad2026.363

Introduction

This study asks whether business students make more responsible decisions when AI-supported analysis includes explicit risk modeling and structured reflection. Generative artificial intelligence has rapidly entered higher education as a tool for explanation, writing support, feedback, ideation, and analysis. Large language models can expand access to information, increase task efficiency, and support

personalized learning processes, but their educational impact remains contingent on how they are integrated into instructional design rather than on the availability of the tool alone [1]. Systematic reviews of artificial intelligence in higher education have also noted that many applications remain technology driven, with insufficient attention to educators, pedagogy, assessment, and the learning processes through which students transform AI-supported activity into durable competence [2], [3]. Recent experimental evidence from secondary-school mathematics also shows that learning consequences depend on how AI support is designed, with pedagogical safeguards mitigating the disadvantages of unrestricted assistance [4].

This issue is particularly relevant in business education. Managerial decision-making requires more than rapid information production or fluent analytical narratives. Students must learn to reason under uncertainty, examine assumptions, interpret risk, evaluate stakeholder consequences, and defend decisions that remain responsible when outcomes are ambiguous. Recent work on generative AI and critical thinking suggests that AI can both support and obscure higher-order cognition, depending on whether students are required to evaluate, challenge, and justify the outputs they receive [5]. Evidence from business schools further suggests that unguided generative AI use may produce uneven effects across students and may alter the distribution of cognitive effort, especially when students use AI-generated outputs as substitutes for analytical monitoring rather than as objects for critical evaluation [6].

The present study is grounded in the view that AI should complement human judgment rather than function as an oracle or autonomous decision-maker. Human-AI collaboration research suggests that hybrid systems are most valuable when AI contributes computational reach and pattern generation while humans retain interpretive responsibility, contextual awareness, and accountability [7]. Human-factors research also cautions that automation can produce misuse, disuse, overreliance, and complacency when users fail to monitor outputs or accept recommendations without sufficient scrutiny [8], [9]. In business education, the instructional challenge is therefore not simply to permit AI use, but to preserve active monitoring and human responsibility within the decision process.

One way to address this challenge is to introduce structured cognitive friction into the human-AI decision loop. In this study, cognitive friction is not treated as an obstacle to learning, but as a deliberate pedagogical mechanism that slows down automatic acceptance, requires validation of assumptions, and redirects attention toward uncertainty, ethics, and consequences. Fan et al. reported stronger essay improvement with ChatGPT but no significant between-group differences in knowledge gain or transfer [10]. This distinction motivates routines that require students to inspect AI-supported reasoning. Monte Carlo simulation serves this purpose by transforming single-point estimates and deterministic recommendations into probabilistic risk maps. This is especially relevant for business education because stochastic simulation can help students understand variability, sensitivity, and risk exposure in decisions involving demand, cost, implementation, and stakeholder impact [11]. More broadly, decision research has long shown that judgments under risk are shaped by biases, reference points, and perceptions of loss and gain, making explicit probabilistic reasoning essential for responsible managerial choice [12].

The second mechanism is a temporal-emotional decision matrix developed for this study. The matrix requires students to examine four dimensions before issuing an executive recommendation: present emotional pressures, present rational arguments, future emotional consequences, and future rational consequences. This structure builds on experiential learning, where competence develops through action, reflection, conceptualization, and renewed experimentation [13]. It also aligns with learning-oriented assessment, where feedback, criteria, and performance tasks are designed not only to measure learning but to shape it [14]. From a metacognitive perspective, the matrix functions as a cognitive forcing tool because it requires students to monitor their own reasoning, identify emotional and cognitive biases, and regulate their decision process before acting [15]. In addition, its stakeholder orientation reflects the idea that business decisions should be evaluated not only by economic feasibility, but also by their consequences for actors affected by the organization [16].

To capture this integrated capability, the study proposes the construct of Responsible Augmented Business Judgment (RABJ). RABJ refers to the learner's capacity to use generative AI as an analytical support system while retaining human accountability for validating assumptions, modeling uncertainty, evaluating stakeholder consequences, regulating cognitive and emotional biases, and defending ethically responsible managerial decisions under ambiguity. This construct differs from AI literacy because it focuses not only on tool use, but on accountable managerial application. It differs from general critical thinking because it is situated in business decisions involving uncertainty, AI-generated advice, and stakeholder consequences. It also differs from business analytics because it treats quantitative analysis as one component of responsible judgment rather than as the endpoint of decision-making.

This study evaluates whether a structured instructional architecture that augments a common AI-supported managerial task with Monte Carlo risk modeling and temporal-emotional reflection is associated with stronger RABJ-related outcomes than AI-supported traditional strategic analysis. RQ1 asks whether the full architecture is associated with stronger gains in the overall RABJ-related composite than the active comparison condition. RQ2 asks which prespecified dimensions show the strongest baseline-adjusted differences between conditions. RQ3 asks whether team performance and qualitative evidence converge with the survey findings. The study makes three contributions: conceptually, it proposes RABJ as a provisional multidimensional applied capability; pedagogically, it operationalizes productive cognitive friction within the human AI decision loop; and empirically, it evaluates the added value of structured uncertainty modeling and reflective accountability against an active AI-supported comparison.

Background and related work

This study builds on the premise that the educational value of generative AI depends on how students are required to interact with it. In business education, AI can support rapid information synthesis, scenario generation, and analytical drafting, but these affordances may also reduce the need for

students to formulate assumptions, test alternatives, and monitor the quality of their own reasoning. For this reason, the proposed intervention does not treat AI access as the instructional innovation itself. Instead, AI is embedded within a structured decision architecture designed to preserve human accountability while expanding analytical capacity.

The first theoretical mechanism is human-AI complementarity. From this perspective, AI contributes computational fluency, rapid synthesis, and alternative generation, while the learner remains responsible for interpretation, contextual judgment, and final decision accountability [7]. This distinction is central to RABJ because the target outcome is not AI literacy alone, but the ability to use AI outputs as objects of evaluation. Students must determine whether AI-generated claims are plausible, incomplete, biased, or insufficiently sensitive to risk and stakeholder implications. This design directly addresses concerns about automation bias and over reliance, where users may accept automated recommendations without adequate monitoring or independent verification [8], [9].

The second mechanism is probabilistic grounding through Monte Carlo simulation. Managerial recommendations generated by AI often appear as coherent narratives, but coherence does not guarantee robustness under uncertainty. Monte Carlo simulation introduces stochastic variation into the decision process by requiring students to model input uncertainty, observe outcome distributions, and interpret risk exposure rather than relying on single-point estimates [11]. In the proposed architecture, Monte Carlo simulation functions as a corrective to deterministic reasoning. It transforms the decision task from “Which recommendation sounds best?” into “Which recommendation remains defensible across plausible scenarios?”

The third mechanism is temporal-emotional metacognition. The temporal-emotional decision matrix requires students to pause before issuing a final recommendation and examine their decision from four angles: present rational arguments, present emotional pressures, future rational consequences, and future emotional consequences. This structure translates metacognitive monitoring into an applied business decision routine, motivated by concerns about dependence on AI during self-regulated learning [10]. Students are required to identify assumptions, emotional triggers, stakeholder consequences, and possible future regret before finalizing the executive recommendation. In this sense, the matrix operates as a cognitive forcing function because it intentionally interrupts premature closure and redirects attention toward reflective judgment [15].

The fourth mechanism is faculty scaffolding and rubric-based accountability. The intervention was not designed as unsupervised AI use. Faculty guidance, structured prompts, performance criteria, and post-task reflection were used to calibrate standards of evidence and decision quality. This is consistent with learning-oriented assessment, where assessment criteria and feedback are part of the learning process rather than only a measurement device [14]. The rubric therefore evaluated not only whether a student reached a recommendation, but whether the decision incorporated uncertainty, AI critique, stakeholder reasoning, ethical defensibility, and future-oriented reflection.

Fig. 1 summarizes the instructional architecture. Role based managerial scenarios and a generative AI tool serve as learning inputs. Monte Carlo modeling, the temporal-emotional decision matrix, and faculty-mediated rubric design operate as instructional mechanisms. The intended outcome is Responsible Augmented Business Judgment, measured through survey outcomes, expert-validated instruments, rubric-based performance assessment, and qualitative coding of student reflections.

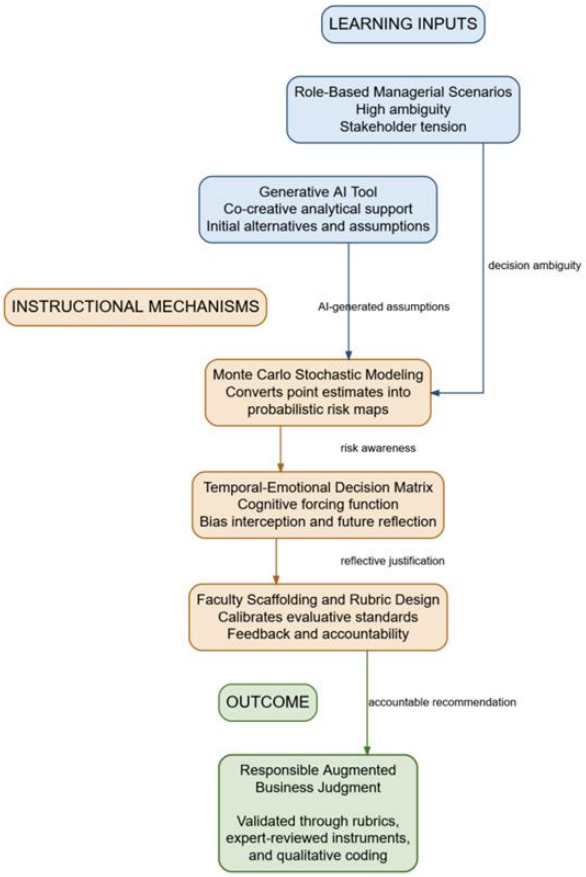


Fig.1 Pedagogical architecture for Responsible Augmented Business Judgment.

Pedagogical Architecture and Methods

Research Design and Participants

A mixed-methods quasi-experimental classroom study was conducted with 120 undergraduate students at a private university in Mexico. Students were nested within 24 teams of five, which were nested within four intact academic sections; teams were the units of condition allocation. Condition allocation was balanced within section blocks, with three teams in the active comparison and three in the full intervention in each section, yielding 12 teams and 60 students per condition. This blocked classroom structure preserved the instructional ecology while reducing section-level imbalance [17]. Both conditions used the same managerial case, role structure, faculty guidance, executive deliverable, and common performance criteria.

All 120 students completed the pre- and post-intervention survey measures. Team-level performance evidence consisted of 24 final recommendations, and anonymized individual reflections were retained for descriptive triangulation [18]. Participation in the research component was voluntary and based on written informed consent under institutional procedures for minimal-risk educational research. Students could decline or withdraw; research records were kept confidential, de-identified before analysis,

and reported only in aggregate. Separate formal committee approval was not required under the applicable institutional procedure.

The active comparison was deliberately substantive rather than passive: both conditions used generative AI and produced an executive recommendation. Across sections, the sequence followed the same broad order—pretest, common case framing and AI-supported diagnosis, condition-specific analysis, final recommendation, posttest, and individual reflection. The experimental contrast therefore concerned the structured uncertainty and reflection mechanisms added in the full intervention rather than simple exposure to AI.

Instructional Conditions

Both conditions worked with the same managerial case and role structure. In the active comparison, students used generative AI to develop a strategic and financial diagnosis of the case and formulate an executive recommendation. Their work focused on examining the plausibility of assumptions, considering stakeholder implications, and justifying the recommendation through conventional strategic and financial reasoning. This condition included faculty guidance, the common assessment criteria, and individual reflection, but did not require stochastic simulation or the structured temporal-emotional matrix. The full intervention retained these elements and added a structured Monte Carlo phase in which key uncertain assumptions were translated into stochastic inputs and interpreted through outcome distributions, downside exposure, sensitivity, and robustness rather than single-point estimates [11]. Students then completed the investigator-developed temporal-emotional decision matrix, which required them to reconsider the recommendation across present emotional pressures, present rational evidence, future emotional consequences, and future rational consequences. The matrix functioned as a forcing routine before the final recommendation, consistent with the intervention's metacognitive and stakeholder rationale [15], [16]. The comparison therefore evaluates Monte Carlo modeling and structured temporal-emotional reflection as a combined addition to the shared AI-supported task; the design does not separate their individual contributions or interaction.

Measures

The study used three outcome evidence sources and one measurement-quality source. First, an investigator-developed 16-item pre-post survey measured four prespecified domains: probabilistic reasoning and risk mitigation (5 items), ethical and stakeholder reasoning (4), critical AI use (3), and metacognition with emotional foresight (4).

Items used a five-point Likert scale. Domain scores were calculated as the mean of their constituent items, and the overall 16-item mean was treated as a provisional composite summary rather than evidence of a unidimensional latent construct. Second, a 10-criterion neutral performance rubric assessed the common team recommendation without requiring use of intervention-specific tools; criteria covered problem framing, evidence use, assumption transparency, uncertainty, alternative comparison, risk mitigation, stakeholder impact, ethical accountability, temporal consequences, and integrated recommendation quality.

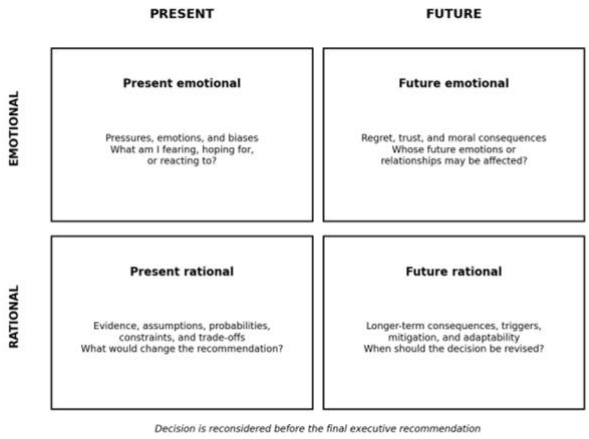


Fig. 2. Temporal-emotional decision matrix used in the full intervention. Students reconsidered the decision across present/future and emotional/rational perspectives before finalizing the executive recommendation.

Two evaluators scored each artifact, and the two-rater mean was used for performance comparison. Third, anonymized reflections were coded for uncertainty reasoning, critical AI use, stakeholder reasoning, ethical reasoning, emotional foresight, and metacognition; only prompts common to both conditions were eligible for between condition comparison. Separately, seven faculty judges reviewed the survey items and rubric criteria for relevance and clarity to provide content-validity evidence [19], [20].

Data Analysis

The analysis followed a triangulated mixed-methods strategy [18]. Descriptive statistics were computed for pre-, post-, and gain scores, and baseline equivalence was examined before intervention effects were tested. Baseline-adjusted ANCOVA was the primary survey analysis, with post-intervention scores as dependent variables, condition as the predictor, and the corresponding pre-intervention score as covariate. Survey models used individual students as the analysis units and did not adjust standard errors for dependence within teams or sections. Gain-score comparisons were retained as a complementary sensitivity description, and Hedges' g summarized between-condition mean gain differences standardized by the pooled within-condition standard deviation of individual gains, with a small-sample correction [21]. Because the four domain outcomes were prespecified and correlated, interpretation emphasizes effect magnitudes and 95% confidence intervals alongside p-values rather than treating each significance test as an independent discovery. Content validity was examined using S-CVI/Ave and Aiken's V [19], [20]. Team performance was interpreted at the team level; ICC(2,k), a two-way random effects, Fig. 2. Temporal-emotional decision matrix used in the full intervention. Students reconsidered the decision across present/future and emotional/rational perspectives before finalizing the executive recommendation. C. Measures The study used three outcome evidence sources and one measurement-quality source. First, an investigator-developed 16 item pre-post survey measured four prespecified domains: probabilistic reasoning and risk mitigation (5 items), ethical and stakeholder reasoning (4), critical AI use (3), and metacognition with emotional foresight (4). Items used a five-point Likert scale. Domain scores were calculated as the mean of their constituent items, and the overall 16-item mean was treated as a provisional composite summary rather than evidence of a unidimensional latent construct. Second, a 10-

criterion neutral performance rubric assessed the common team recommendation without requiring use of intervention-specific tools; criteria covered problem framing, evidence use, assumption transparency, uncertainty, alternative comparison, risk mitigation, stakeholder impact, ethical accountability, temporal consequences, and integrated recommendation quality. Two evaluators scored each artifact, and the two-rater mean was used for performance comparison. Third, anonymized reflections were coded for uncertainty reasoning, critical AI use, stakeholder reasoning, ethical reasoning, emotional foresight, and metacognition; only prompts common to both conditions were eligible for between condition comparison. Separately, seven faculty judges reviewed the survey items and rubric criteria for relevance and clarity to provide content-validity evidence [19], [20]. D. Data Analysis The analysis followed a triangulated mixed-methods strategy [18]. Descriptive statistics were computed for pre-, post-, and gain scores, and baseline equivalence was examined before intervention effects were tested. Baseline-adjusted ANCOVA was the primary survey analysis, with post-intervention scores as dependent variables, condition as the predictor, and the corresponding pre-intervention score as covariate. Survey models used individual students as the analysis units and did not adjust standard errors for dependence within teams or sections. Gain-score comparisons were retained as a complementary sensitivity description, and Hedges' g summarized between condition mean gain differences standardized by the pooled within-condition standard deviation of individual gains, with a absolute-agreement, average-measures coefficient, quantified rater agreement and was not treated as a test of the condition difference [22]. This rater-agreement coefficient does not quantify within-team dependence among survey responses. Qualitative evidence was used for descriptive triangulation rather than an independent causal test, and only aggregate coded patterns are reported. Given the intact section and team ecology, student-level ANCOVA estimates are interpreted as baseline adjusted classroom associations rather than component-specific causal effects.

Table 1. Instructional Conditions and Evidence Sources

Component	AC	FI	Implementation / evidence
AI-supported strategic analysis	Yes	Yes	Classroom task
Executive recommendation	Yes	Yes	Final team output
Role-based managerial scenario	Yes	Yes	Common case activity
Monte Carlo risk modeling	No	Yes	Simulation output
Temporal-emotional matrix	No	Yes	Reflection protocol
Faculty guidance	Yes	Yes	Common instructional support
Rubric-based performance assessment	Yes	Yes	Two-rater team scores
Individual reflection	Yes	Yes	Common + condition-specific prompts
Expert content validation	NA	NA	Seven faculty judges
Qualitative reflection coding	Yes	Yes	Descriptive triangulation

a. AC = active comparison; FI = full intervention; NA = not applicable. Both conditions shared the case, role structure, AI access, guidance, and assessment; FI added Monte Carlo and temporal-emotional reflection.

Results and findings

Data Integrity and Measurement Evidence

The analytic dataset contained 120 complete student records, with 60 students in each condition and no missing pre- or post-survey values. Team-level performance data comprised 24 final recommendations, 12 per condition. No duplicate student identifiers or out-of-range Likert responses were detected.

Expert review supported the content validity of the measurement system. Across the 16 survey items and 10 neutral rubric criteria, S CVI/Ave was .984 for relevance and .989 for clarity; mean Aiken's V was .870 and .828, respectively. Internal consistency was interpreted cautiously because RABJ was conceptualized as a multidimensional applied capability rather than a single reflective trait. The overall post-intervention composite showed acceptable reliability, $\alpha = .786$, whereas pre-intervention reliability was lower, $\alpha = .400$. The baseline coefficient therefore limits interpretation of the overall pretest composite and increases the importance of domain-level and non-survey evidence.

Pre and Post Survey Outcomes

Survey outcomes were analyzed using the four prespecified domain means and the overall 16-item composite described above. Results are reported using both gain comparisons and baseline-adjusted contrasts so that descriptive change and covariate-adjusted differences can be distinguished.

Table 2 summarizes the main survey outcomes. Both conditions improved from pre to post measurement, but gains were consistently larger in the full intervention. The overall composite increased by 0.407 in the active comparison and by 0.898 in the full intervention, a gain difference of 0.491. The largest gain difference appeared in metacognition and emotional foresight, followed by probabilistic reasoning and risk mitigation.

Baseline-adjusted ANCOVA was used as the primary survey analysis because small pre-intervention differences were present in some outcomes. The full intervention was associated with higher adjusted post-intervention scores across all prespecified domains. The adjusted difference on the overall composite was $B = 0.557$, 95% CI [0.482, 0.632], $p < .001$. The largest adjusted difference was observed for metacognition and emotional foresight, $B = 0.793$, 95% CI [0.652, 0.934], $p < .001$; probabilistic reasoning and risk mitigation also showed a strong adjusted difference, $B = 0.569$, 95% CI [0.426, 0.712], $p < .001$. Fig. 3 visualizes the baseline-adjusted condition contrasts across the five survey outcomes.

Performance-Based Evidence

Team-level performance evidence converged directionally with the survey results. Across 24 final recommendations, the full intervention obtained a higher two-rater mean rubric score than the active comparison (3.47 vs. 2.93). Agreement between evaluators was high for the overall rubric score, ICC(2,k) = .954. Because ICC quantifies rater agreement rather than the significance of the condition contrast, the between-condition rubric difference is interpreted here as

convergent performance evidence rather than as a separate inferential confirmation.

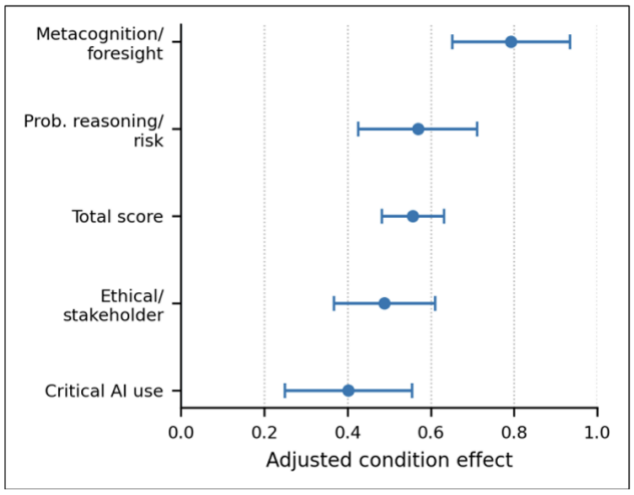


Fig. 3. Baseline-adjusted ANCOVA effects favoring the full intervention across RABJ outcomes. Error bars represent 95% confidence intervals.

Table 2. Main Survey Outcomes by Condition

Outcome	AC Δ	FI Δ	g	Adj. B [95% CI]
Prob. reasoning / risk	0.450	0.913	1.040	0.569 [0.426, 0.712]
Ethical / stakeholder	0.388	0.758	0.751	0.489 [0.367, 0.610]
Critical AI use	0.550	0.828	0.470	0.402 [0.249, 0.555]
Metacog. / foresight	0.267	1.071	1.481	0.793 [0.652, 0.934]
Overall composite	0.407	0.898	1.951	0.557 [0.482, 0.632]

a. AC = active comparison; FI = full intervention; Δ = pre-post mean change; g = Hedges' g for between-condition gain differences; Adj. B = unstandardized baseline-adjusted posttest contrast. b. All adjusted effects were significant at $p < .001$. g and Adj. B represent different estimands and are not expected to be numerically equivalent.

Qualitative Triangulation and Evidence Integration

Qualitative coding of anonymized reflections showed stronger full-intervention patterns of uncertainty awareness, critical AI use, stakeholder sensitivity, ethical reasoning, emotional foresight, and metacognitive reflection. To reduce prompt-induced bias, only prompts common to both conditions were used for direct condition comparisons; intervention-specific reflection prompts were treated as descriptive process evidence only. The qualitative component therefore contextualizes and corroborates the quantitative pattern without functioning as an independent causal test [18].

Table 3 summarizes the inferential role of each evidence stream. The distinction is important because measurement validation, student self-report, team performance, and qualitative triangulation answer different questions and should not be treated as interchangeable evidence.

Table 3. Evidence Triangulation and Inferential Role

Evidence	Observed pattern	Interpretive role
Survey	FI larger gains; overall Adj. B = .557; all adjusted domain effects $p < .001$.	Primary student-level evidence; total composite interpreted cautiously.
Team rubric	FI = 3.47 vs. AC = 2.93; ICC(2,k) = .954.	Convergent team-level performance evidence; condition contrast descriptive.
Reflections	Stronger FI patterns across six coded themes.	Descriptive triangulation; only common prompts compared.
Expert review	S-CVI/Ave = .984/.989; Aiken's V = .870/.828.	Measurement quality only; not an intervention outcome.

Discussion

The findings suggest that generative AI is more educationally valuable when students are required to evaluate, constrain, and justify its outputs rather than use it primarily to accelerate analytical production. Because both conditions used AI, the observed pattern is more consistent with an instructional architecture effect than with a simple technology-access effect. The full intervention was associated with stronger outcomes across the prespecified survey domains and showed convergent team-performance and reflection patterns, supporting the central argument that the relevant pedagogical question is how the human-AI decision loop is structured.

The large overall gain effect (g = 1.951) should be interpreted in relation to its denominator: the pooled within condition standard deviation of individual gain scores [21]. The unstandardized gain difference was 0.491 points on the five point scale; a relatively small dispersion of gains yields a large standardized contrast. Averaging 16 items can reduce dispersion, so the composite effect is not directly comparable with effects for individual domains or differently standardized outcomes. The immediate self-report measure and its close alignment with the intervention may also contribute to the magnitude. These are plausible explanations, not demonstrated mechanisms, and the estimate should not be read as evidence of durable or generalizable causal learning gains.

The strongest adjusted difference appeared in metacognition and emotional foresight. This pattern is theoretically consistent with the temporal-emotional matrix, which required students to pause before issuing a recommendation, identify present pressures, anticipate future consequences, and connect immediate analysis with longer-term ethical and stakeholder implications. The alignment is mechanism-consistent rather than mechanism-proving: because both added components were delivered together, the design cannot attribute the observed differences to either one or test their interaction.

The probabilistic reasoning and risk-mitigation results are likewise consistent with the intended role of Monte Carlo

simulation. AI-generated recommendations can appear coherent while relying on fragile assumptions or deterministic simplifications. Requiring students to examine distributions, downside exposure, sensitivity, and decision robustness shifted the task from producing a plausible answer toward evaluating whether that answer remained defensible across uncertain scenarios, consistent with prior simulation-based decision learning in business contexts [11].

More broadly, the architecture reframes AI-related cognitive efficiency. The pedagogical objective is not to minimize cognitive effort, but to relocate it from information production toward validation, uncertainty interpretation, and accountable justification. This interpretation is consistent with concerns that automation can displace monitoring [8], [9], with evidence that generative AI can redistribute cognitive effort in business school tasks [6], and with metacognitive approaches that deliberately require learners to inspect and regulate their own reasoning [15].

Performance-based and qualitative evidence strengthened—but did not replace—the survey interpretation. The higher team rubric mean indicates that the pattern was not limited to self-report, while high interrater reliability supports scoring stability. Reflection coding provided explanatory convergence across uncertainty awareness, AI critique, stakeholder sensitivity, ethical reasoning, emotional foresight, and metacognition. At the same time, Table 3 makes the inferential boundaries explicit: rubric means are descriptive between conditions, qualitative evidence is contextual, and expert review supports measurement quality rather than intervention effectiveness.

Conceptually, RABJ should therefore be treated at this stage as a provisional multidimensional applied capability rather than a fully validated psychometric trait. The domain structure, the low pre-intervention alpha of the overall composite, and the use of multiple evidence streams all argue against relying on a single total score. The stronger interpretation rests on convergence across prespecified domains, externally rated team performance, and reflection evidence, consistent with the mixed-methods logic of complementarity and triangulation [18].

The study also has practical implications for business education. AI integration should not be limited to allowing generative tools in assignments; instructors can design sequences that require students to disclose assumptions, model uncertainty, challenge AI outputs, consider affected stakeholders, and justify decisions using explicit performance criteria. The architecture is intentionally modular: the same logic can be adapted to international business, finance, operations, entrepreneurship, sustainability, and strategy courses in which students must defend decisions under uncertainty.

Several limitations qualify these findings. The quasi experimental, single-institution design limits causal and contextual generalization, and outcomes capture immediate rather than durable change. Although condition allocation was balanced within sections, students shared team activities. The student-level ANCOVA did not adjust for this dependence; its confidence intervals may therefore be too narrow and p-values too small if residual within-team or within-section correlation is positive. The intervention bundled Monte Carlo modeling with temporal-emotional reflection and related scaffolding, so component-specific contributions cannot be isolated. The low pre-intervention reliability of the overall survey composite warrants caution, and RABJ remains a provisional framework rather than a fully validated latent construct. Team rubric differences are descriptive between conditions, while qualitative evidence is non-inferential and constrained to comparable prompts. Replication with cluster-aware models, independent samples, delayed retention measures, and component-level tests is therefore needed.

Conclusion

This study advances a practical model for integrating generative AI into business education without reducing learning to faster output production. In a common AI-supported managerial task, adding Monte Carlo risk modeling and temporal-emotional reflection was associated with stronger baseline-adjusted survey outcomes and convergent-team performance and reflection evidence. The central contribution is therefore not AI as a tool in itself, but an instructional architecture that keeps students responsible for assumptions, uncertainty, stakeholder consequences, and the defensibility of final decisions. Future research should replicate the architecture across institutions and disciplines, use cluster-aware and longitudinal designs, and isolate the contribution of individual mechanisms. RABJ is best viewed here as a provisional multidimensional capability that provides a testable framework for designing and evaluating accountable human-AI judgment in management education.

Acknowledgment

The authors gratefully acknowledge the financial support provided by the International Business Department, School of Business, Tecnológico de Monterrey, Mexico, in support of this work. The authors also express sincere appreciation to the participating students for their engagement, collaboration, and thoughtful reflections, which contributed meaningfully to this research.

References

[1] E. Kasneci et al., “ChatGPT for good? On opportunities and challenges of large language models for education,” Learning and Individual Differences, vol. 103,. Art. 102274, 2023, doi: 10.1016/j.lindif.2023.102274

[2] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, “Systematic review of research on artificial intelligence applications in higher education: Where are the educators?,” International Journal of Educational Technology in Higher Education, vol. 16, Art. 39, 2019, doi: 10.1186/s41239-019-0171-0.

[3] H. Crompton and D. Burke, “Artificial intelligence in higher education: The state of the field,” International Journal of Educational Technology in Higher Education, vol. 20, Art. 22, 2023, doi: 10.1186/s41239-023-003928.

[4] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman, “Generative AI without guardrails can harm learning: Evidence from high school mathematics,” Proceedings of the National Academy of Sciences, vol. 122, no. 26, Art. e2422633122, 2025, doi: 10.1073/pnas.2422633122.

[5] C. Gonsalves, “Generative AI’s impact on critical thinking: Revisiting Bloom’s taxonomy,” Journal of Marketing Education, early access, 2024, doi: 10.1177/02734753241305980.

[6] C. Bergenholtz, O. Vuculescu, F. Günzel-Jensen, and L. Frederiksen, “Leveling up or leveling down? The impact of generative

AI on student performance in business schools," Academy of Management Learning & Education, early access, 2026, doi: 10.5465/amle.2025.0029.
[7] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister, "Hybrid intelligence," Business & Information Systems Engineering, vol. 61, no. 5, pp. 637–643, 2019, doi: 10.1007/s12599-019-00595-2.
[8] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," Human Factors, vol. 39, no. 2, pp. 230–253, 1997, doi: 10.1518/001872097778543886.
[9] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," Human Factors, vol. 52, no. 3, pp. 381–410, 2010, doi: 10.1177/0018720810376055.
[10] Y. Fan et al., "Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance," British Journal of Educational Technology, vol. 56, no. 2, pp. 489-530, 2025, doi: 10.1111/bjet.13544.
[11] A. Pritchard, D. F. Taylor, and M. Belford, "Teaching data-driven decision making for inventory analysis with Monte Carlo simulation," Decision Sciences Journal of Innovative Education, early access, 2024, doi: 10.1111/dsji.12328.
[12] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," Econometrica, vol. 47, no. 2, pp. 263–291, 1979, doi: 10.2307/1914185.
[13] D. A. Kolb, Experiential Learning: Experience as the Source of Learning and Development. Englewood Cliffs, NJ, USA: Prentice Hall, 1984.
[14] D. Carless, "Learning-oriented assessment: Conceptual bases and practical implications," Innovations in Education and Teaching International, vol. 44, no. 1, pp. 57–66, 2007, doi: 10.1080/14703290601081332.
[15] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry," American Psychologist, vol. 34, no. 10, pp. 906–911, 1979, doi: 10.1037/0003-066X.34.10.906.
[16] R. E. Freeman, Strategic Management: A Stakeholder Approach. Boston, MA, USA: Pitman, 1984.
[17] W. R. Shadish, T. D. Cook, and D. T. Campbell, Experimental and Quasi Experimental Designs for Generalized Causal Inference. Boston, MA, USA: Houghton Mifflin, 2002.
[18] J. W. Creswell and V. L. Plano Clark, Designing and Conducting Mixed Methods Research, 3rd ed. Thousand Oaks, CA, USA: SAGE, 2018.
[19] D. F. Polit, C. T. Beck, and S. V. Owen, "Is the CVI an acceptable indicator of content validity? Appraisal and recommendations," Research in Nursing & Health, vol. 30, no. 4, pp. 459–467, 2007, doi: 10.1002/nur.20199.
[20] L. R. Aiken, "Three coefficients for analyzing the reliability and validity of ratings," Educational and Psychological Measurement, vol. 45, no. 1, pp. 131–142, 1985, doi: 10.1177/0013164485451012.
[21] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs," Frontiers in Psychology, vol. 4, Art. 863, 2013, doi: 10.3389/fpsyg.2013.00863.
[22] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," Journal of Chiropractic Medicine, vol. 15, no. 2, pp. 155–163, 2016, doi: 10.1016/j.jcm.2016.02.012.