ORIGINAL ARTICLE

Impacts of an Artificial Intelligence Tutor in Foundation Physics: Cognitive Load, Confidence, Interest, Calibration, and Performance

Ali Abu-Nada[1]*, Rami Shariah[2], Saud Al-Dajah[1]

[1]Sharjah Maritime Academy, Sharjah, United Arab Emirates, [2]Department of Support and School Development, Ministry of Education, Dubai, United Arab Emirates

***Corresponding Author:** ali.abunada@sma.ac.ae

ABSTRACT

This study investigates whether a GPT-based tutor can improve learning and self-monitoring in introductory physics. Two parallel sections studied Newton's Second Law: An experimental class used the tutor artificial intelligence (AI) and a control class used a textbook (CO). Students then took a quiz with per-question confidence ratings and completed brief questionnaires on extraneous cognitive load, intrinsic cognitive load, self-efficacy, situational interest (SI), and affect (EMO). We examined quiz performance, confidence patterns, and metacognitive calibration, how closely confidence matches accuracy. AI students scored higher and showed better calibration: they tended to be confident when correct and less confident when wrong. CO students reported higher overall confidence but showed weaker calibration. Questionnaire reliability (Cronbach's α) was generally higher for AI, with SI and EMO most consistent. Overall, the GPT-based tutor improved performance, supported motivation, and strengthened metacognitive judgment in a foundation-level physics setting.

KEY WORDS: Artificial intelligence tutor; Physics education; Metacognition; Confidence calibration; Cognitive load; ChatGPT

INTRODUCTION

Teaching introductory physics to 1st-year university students remains challenging, many struggle to construct robust conceptual models of abstract ideas (e.g., force, energy, fields) and to transfer those models to novel contexts and quantitative problem solving (Hestenes et al., 1992; Hake, 1998; Mazur, 1997; McDermott and Shaffer, 2002; Ambrose et al., 2010; Bransford et al., 2000). Beyond content coverage, effective instruction must also cultivate metacognitive skills, learners' capacity to plan, monitor, and evaluate their own understanding and strategies (Flavell, 1979; Schraw and Dennison, 1994; Zimmerman, 2002). When these skills are weak, students often misjudge their knowledge, either showing overconfidence or withholding confidence despite correct reasoning. This mismatch between confidence and correctness, calibration, matters because well-calibrated learners allocate study time more effectively, surface misconceptions earlier, and solve problems more efficiently than poorly calibrated peers (Ambrose et al., 2010; Bransford et al., 2000).

Recent advances in artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT (Jia et al., 2023), have introduced new possibilities for addressing these challenges. Unlike static textbooks or scripted tutorials, AI-driven tutors can provide dynamic, conversational, and context-sensitive explanations tailored to individual learner needs. They can offer step-by-step guidance, respond to follow-up questions, and generate targeted practice problems, capabilities that can enhance both cognitive and metacognitive dimensions of learning. Importantly, these systems can also elicit real-time self-assessment (e.g., through confidence ratings) and provide feedback aligned with students' levels of certainty, creating opportunities for deliberate calibration training.

A rapidly expanding body of research has explored how LLM-based chatbots such as ChatGPT can support teaching and learning across disciplines. Their interactive dialogue capabilities allow them to deliver immediate feedback, scaffold reasoning, and adapt explanations in ways that traditional static materials cannot (Kasneci et al., 2023; Chan, 2023; Bitzenbauer, 2023). Students frequently report that AI tutors increase accessibility and motivation, particularly in large classes where individualized instructor feedback is limited (Sun and Zhou, 2022). Moreover, they can reduce extraneous cognitive load (ECL) by breaking complex problems into manageable steps and promoting deeper learning through guided reflection and elaboration (Graesser et al., 2016; McNamara et al., 2013). Importantly, chatbots have also been shown to influence metacognitive processes: By prompting learners to articulate reasoning, assess their confidence, and reflect on mistakes, they can improve calibration and self-regulation skills that are essential for long-term learning success (Schraw and Dennison, 1994; Zimmerman, 2002).

However, the literature also highlights several challenges and risks. Overreliance on AI-generated explanations can lead students to accept incorrect responses uncritically, while the anthropomorphic framing of conversational agents may cause them to overtrust the system or attribute human-like understanding to it (Bewersdorff et al., 2023; Pizzi et al., 2023). Furthermore, questions remain about how these tools impact different aspects of learning – from conceptual understanding to motivation and self-efficacy (SE) – and how their effects vary across disciplines and student populations.

Within physics education specifically, the integration of AI tutors is still at an early stage, and several open questions remain. One is whether AI support leads to measurable improvements in student performance compared with conventional textbook-based study. Another is whether AI-mediated interactions influence metacognitive outcomes such as confidence calibration, guessing behavior, and the coupling between confidence and accuracy. Understanding these dynamics is crucial because overconfidence, particularly when paired with low accuracy, can entrench misconceptions, whereas underconfidence can discourage persistence even when students are on the right track.

The present study addresses these gaps by conducting a controlled classroom experiment in two parallel sections of a foundational physics course. One section engaged with a custom-configured AI tutor designed to explain, coach, and generate practice problems, while the other followed a traditional textbook-based approach. Both groups studied the same topic (Newton's Second Law) and completed identical assessments, including a 10-item written quiz with per-item confidence ratings and short validated questionnaires measuring cognitive load, SE, situational interest (SI), and affect. By analyzing performance, confidence profiles, calibration metrics, and internal consistency of questionnaire scales, we aim to provide a transparent, reproducible comparison of AI-assisted and traditional instruction.

Our central research questions are (i) whether a conversational AI tutor improves student performance in foundational physics relative to textbook-based instruction; (ii) whether AI support affects metacognitive outcomes, particularly confidence calibration and the alignment between confidence and accuracy; and (iii) how AI tutoring influences cognitive load, SE, and SI compared with traditional methods.

By addressing these questions, this study contributes to the growing literature on AI in STEM education and provides evidence-based insights for instructors and curriculum designers seeking to integrate AI tools effectively. Beyond immediate learning outcomes, our findings also inform the broader conversation on how generative AI can be leveraged to cultivate metacognitive skills-skills that are essential not only for mastering physics but also for lifelong learning in science and engineering.

METHODS

Participants, Materials, and Procedure

The study was conducted at Sharjah Maritime Academy, a higher-education institution in the United Arab Emirates specializing in maritime sciences, engineering, and applied technology. Two intact sections of a 1st-year introductory physics course participated. The focal topic was Newton's Second Law, which had not yet been taught in class at the time of data collection.

Students took part in a guided pre-instruction learning session, during which they studied materials specific to their assigned group. The experimental section (AI) received a GPT-generated handout aligned with the course syllabus, including structured explanations, worked examples, and guided problem-solving prompts. The control section (CO) received a matched handout compiled from the required textbook, Physics of Everyday Phenomena (10th ed.) by W. Thomas Griffith and Juliet Brosing (Griffith and Brosing, 2023). Students were given 30 min for self-study with their assigned handout before assessment activities.

The two handouts were carefully matched for topic coverage, length, notation, and example complexity to ensure content equivalence across conditions. Both sections were taught by the same instructor, followed the same syllabus, and shared identical contact hours and assessment conditions. Representative excerpts of the handouts, a sample of the 10-item quiz (with per-item confidence ratings), and the post-assessment questionnaire items are provided in Appendix A.

The post-quiz questionnaire consisted of five three-item scales designed to measure distinct dimensions of students' cognitive and emotional learning experiences: ECL, intrinsic cognitive load (ICL), SE, SI, and Affect (EMO). Each scale used a 5-point Likert format (1 = strongly disagree to 5 = strongly agree).

Procedure

The study followed a structured sequence within two class sessions. Students had not previously been taught Newton's Second Law, ensuring that performance reflected the pre-instruction materials. The procedure unfolded as follows:

1. Room setup. Students were seated apart, and condition-specific handouts (AI or textbook) were distributed. Use of personal devices or additional resources was not permitted.
2. Self-study (30 min). Students engaged in independent study using the assigned handout. The instructor announced the remaining time at 15 and 5 min.
3. Questionnaire (5 min). After studying, students completed a brief questionnaire consisting of five validated three-item scales: ECL, ICL, SE, SI, and affect (EMO). Each scale used a 5-point Likert response format (1 = strongly disagree to 5 = strongly agree).
4. Quiz with confidence (15 min). Students then completed a 10-question quiz on Newton's Second Law under standard

classroom conditions. Immediately after answering each question, they rated their confidence on a 5-point scale (1 = very unsure to 5 = very sure).

5. Debrief. Students were thanked for their participation and reminded that the activity did not affect course grades.

Representative excerpts of both handouts, example quiz items, and the full post-assessment questionnaire are provided in Appendix A.

Measures and Derived Variables

All outcome measures in this study were derived from students' quiz responses, confidence ratings, and questionnaire data collected during the session. Our aim was to capture both cognitive and metacognitive performance. By cognitive performance, we mean how well students understand the ideas and apply them to problems (here, Newton's Second Law). By metacognitive performance, we mean how well students monitor their own thinking, for example, judging when they are likely right or wrong and matching their confidence to their accuracy. Below, we describe each variable, how it was computed, and what it represents.

a. Quiz performance. Each quiz consisted of 10 conceptual and applied questions on Newton's Second Law. Quiz responses were marked as either correct (1) or incorrect (0). From these scores, we derived two primary measures:
   - Raw score (0–10): Total number of correct answers per student.
   - Percent score (0–100%): Raw score divided by 10, providing a normalized performance metric for comparison between groups.

b. Confidence ratings. After answering each quiz question, students recorded their confidence in the correctness of their answer on a 5-point Likert scale (1 = very unsure, 5 = very sure). These ratings were used to derive several measures:
   - Average confidence: The mean confidence rating across all 10 questions, reflecting students' overall certainty during the quiz.
   - Guessing behavior: The number and proportion of quiz responses where confidence was low (≤2), used as an indicator of guessing or uncertainty.

c. Metacognitive calibration. Calibration refers to the alignment between students' confidence and the accuracy of their responses. It reflects how well students "know what they know." For each student, we categorized every response into one of four calibration outcomes:
   - High confidence correct (HCC): The number of correct answers given with high confidence (≥4).
   - High-confidence wrong (HCW): The number of incorrect answers given with high confidence (≥4), indicating overconfidence.
   - Low confidence correct (LCC): The number of correct answers given with low confidence (≤2), indicating under-confidence.
   - Low-confidence wrong (LCW): The number of incorrect answers given with low confidence (≤2).

These calibration categories provide a fine-grained picture of metacognitive accuracy. A profile with many HCC responses and few HCW responses reflects strong calibration. By contrast, frequent HCW responses indicate overconfidence (being sure but wrong), while frequent LCC responses indicate underconfidence (being right but unsure). The most well-calibrated learners are those who maximize HCC while minimizing both HCW and LCC.

d. Cognitive and affective questionnaire scales. After the 30-min preparation period and before beginning the quiz, students completed a brief questionnaire consisting of five validated three-item scales, each measured on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). The scales were:
   - ECL: Effort spent on presentation issues (format, wording, layout, navigation) instead of the physics content (e.g., confusing instructions, cluttered pages, hard-to-find steps).
   - ICL: How hard the physics idea itself is.
   - SE: How sure students feel they can learn and solve the problems.
   - SI: How much the activity grab attention and curiosity right now.
   - Affect (EMO): How students feel during the task (e.g., calm or anxious).

e. Scale reliability (Cronbach's α). Cronbach's α indicates how well the items on a short questionnaire work together as one scale. If the items all tap the same underlying idea, students who score high on one item also tend to score high on the others and α is higher; if items ask about different things, are unclear, or behave inconsistently, α is lower. The coefficient ranges from 0 to 1 (larger values mean greater internal consistency). A common quick guide is (Cronbach, 1951; Tavakol and Dennick, 2011; Sijtsma, 2009): $\alpha \geq 0.90$ excellent, 0.80–0.89 good, 0.70–0.79 acceptable, 0.60–0.69 marginal, <0.60 low.

With only three items, α often runs lower, so we interpret it more cautiously. We average the three item scores to form a single scale score (e.g., an ECL score). α checks whether that average is meaningful: High α means the items act like one ruler and the average is a reliable summary; low α means the items do not behave like one ruler and the average is noisy, so group comparisons and correlations are less trustworthy.

Example. Consider three SE items. If confident students rate all three items high and less confident students rate all three low, the items "move together" across students and α is high. If confident students give mixed answers across the three items (high on one, low on another), the items do not hang together and α is low. A detailed explanation of how Cronbach's α was calculated in this study, including the item-level analysis and reliability computations, is provided in Appendix B.

Together, these measures offer a complementary view of learning outcomes and processes: quiz scores index conceptual understanding; confidence ratings and calibration indices

reflect the accuracy of metacognitive monitoring; and the questionnaire scales capture cognitive load, self-beliefs, motivation, and emotional states. Representative quiz items and the full questionnaire wording are provided in Appendix A.

RESULTS

Performance Outcomes

Student performance was assessed with a 10-item quiz on Newton's Second Law, scored dichotomously (1 = correct, 0 = incorrect). We summarize four group-level indicators using the notation in Table 1: $\bar{S}$ = mean quiz score (0–10), $\bar{C}$ = mean confidence (1–5), L = count of low-confidence items per student (responses with confidence ≤2, out of 10), and L = guess rate (%).

Performance ($\bar{S}$). The AI-tutor group outperformed the textbook control, with $\bar{S}_{AI}$ = 6.65/10 versus $\bar{S}_{CO}$ = 4.45/10 (Table 1). This sizable difference indicates higher immediate conceptual understanding following the AI-supported preparation under identical topic coverage and study time.

Confidence ($\bar{C}$). Despite scoring higher, the AI group reported lower average confidence ($\bar{C}_{AI}$ = 3.31) than the control ($\bar{C}_{CO}$ = 3.68). Thus, the control students felt more certain overall even though their accuracy was lower, showing that mean confidence alone is not a reliable proxy for performance.

Low-confidence responding (L) and guess rate (G). AI students marked fewer low-confidence items on average than control students (L_{AI} = 3.34/10 vs. $L_{.CO}$ = 3.85/10), and they showed a correspondingly lower guess rate (G_{AI} = 33.5% vs. G_{CO} = 38.5%; Table 1). Together with their higher quiz scores, these patterns indicate that AI-supported learners expressed uncertainty less often and guessed less frequently overall, consistent with stronger metacognitive calibration.

Metacognitive Calibration Patterns

Figure 1 shows how students' answers were distributed across four confidence–accuracy categories: HCC, HCW, LCW, and LCC. These categories are designed to capture the quality of students' metacognitive calibration, that is, how well their confidence matched the accuracy of their answers.

Table 1: Descriptive statistics for AI and (CO) groups

Measure	AI (n=17)	CO (n=20)
$\bar{S}$ (Mean quiz score,/10)	6.65	4.45
$\bar{C}$ (Mean confidence,/5)	3.31	3.68
L (Low-confidence items,/10)	3.34	3.85
G (Guess rate, %)	33.5	38.5

Symbols: $\bar{S}$ = mean quiz score (0–10); $\bar{C}$ = mean confidence (1–5); L=low-confidence question per student (count of items with confidence ≤2, out of 10); G=guess rate (%). AI: Artificial intelligence

For each student, every quiz response (10 total) was classified into one of the four categories based on its correctness and the confidence rating they reported:

- HCC: Correct answer with confidence ≥4 (strong and correct)
- HCW: Incorrect answer with confidence ≥4 (overconfident error)
- LCW: Incorrect answer with confidence ≤2 (low confidence and wrong)
- LCC: Correct answer with confidence ≤2 (underconfident success).

We then counted how many questions for each student fell into each category (out of 10) and averaged these counts across all students in a group. Thus, the y-axis in Figure 1 represents the average number of questions per student (out of 10) in each category.

Students supported by the AI tutor, Figure 1a, showed a highly desirable calibration pattern. The HCC bar dominates, with an average of about 4.94 questions per student answered both confidently and correctly. This result indicates that when AI-assisted students felt sure of their answers, they were usually correct, a sign of accurate metacognitive judgment. Meanwhile, HCW responses were low (≈ 0.24), demonstrating that overconfidence was extremely rare. This combination, high HCC and very low HCW, shows that AI students could reliably judge when they understood the material. The remaining two categories, LCW (≈ 2.24) and LCC (≈ 0.76), are moderate to low, indicating that students recognized their uncertainty when they were wrong and rarely underestimated themselves when correct. Overall, this pattern reflects strong monitoring and self-assessment skills: students were not only more accurate but also more aware of the reliability of their own knowledge.

The textbook control group (CO) (Figure 1b) displayed a very different profile. HCC responses were much lower (≈ 2.70), showing that students were confidently correct far less often. More importantly, HCW responses were the most frequent category (≈ 3.30), indicating widespread overconfidence – students frequently believed they were correct.

When they were not. This overconfidence is a classic sign of weak metacognitive calibration, as students' subjective certainty did not align with their actual performance.

The LCW (≈ 1.25) and LCC (≈ 1.15) categories were present at low to moderate levels but did not offset the overconfidence trend.

Together, these results provide strong evidence that AI tutoring supports not just better content learning but also more accurate self-monitoring. The AI group demonstrated the ideal calibration profile: many confident-correct responses and almost no confident wrong ones. This suggests that GPT-based guidance helped students form more accurate internal judgments about their knowledge. In contrast, control students

frequently expressed confidence even when incorrect, a pattern associated with entrenched misconceptions and less effective self-regulation. These findings highlight one of the most significant educational benefits of AI tutoring not only improving accuracy but also cultivating a deeper awareness of what students know and do not know.

Confidence–Accuracy Relationship

Figure 2 illustrates the relationship between students' average confidence and their quiz scores in the AI and CO groups. Each marker is one student: The x–value is the student's average confidence across the 10 quiz items (1–5) and the y–value is the student's quiz score (0–10). Points may overlap because multiple students can have the same pair ($\bar{C}_i$, S_i); therefore, the number of visible points can be smaller than n.

For student i and question j, let $c_{ij} \in \{1...,5\}$ be the reported confidence and $a_{ij} \in \{0, 1\}$ indicate correctness. We computed two student-level summaries:

$$\bar{C}_i = \frac{1}{10}\sum_{j=1}^{10} c_{ij} \quad \text{(Average confidence)} \tag{1}$$

$$S_i = \sum_{j=1}^{10} a_{ij} \quad \text{(quiz score, 0–10)} \tag{2}$$

We then plotted ($\bar{C}_i$, S_i) for all students in a group, fit an ordinary least–squares line.

To quantify the relationship between confidence and performance, we model the quiz score S as a linear function of average confidence $\bar{C}$ using an ordinary least squares (OLS) regression:

$$S = \beta_0 + \beta_1 \bar{C} \tag{3}$$

Where β_0 (y-intercept) is the predicted quiz score when average confidence is zero, and β_1 represents the expected change in quiz score for a one-unit increase in average confidence.

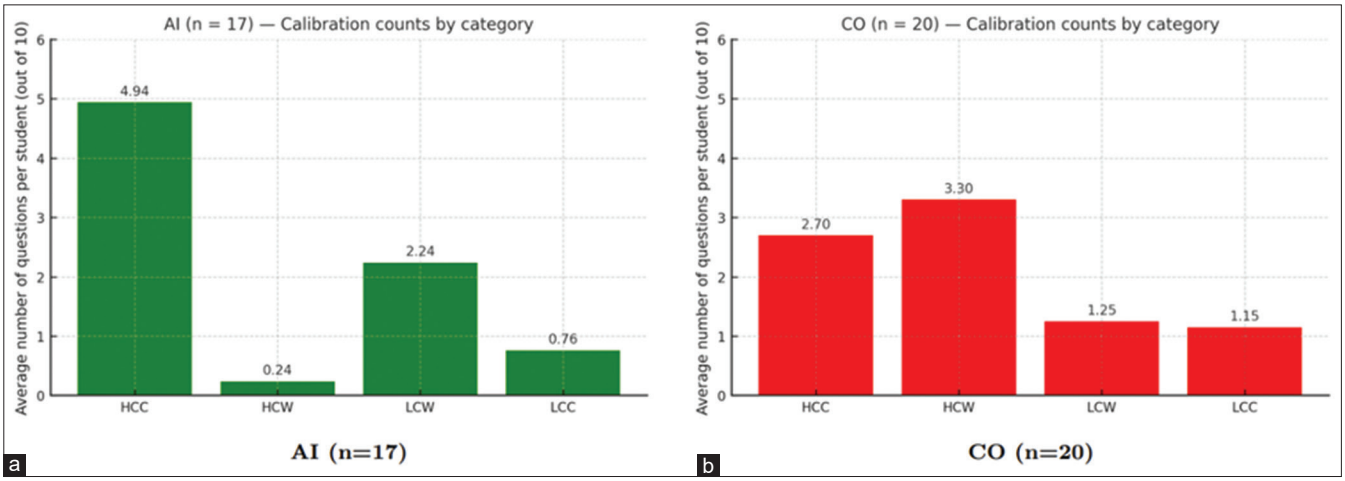


Figure 1: Metacognitive calibration counts per student (out of 10). (a) AI group and (b) control group. Bars show high-confidence correct, high-confidence wrong, low-confidence correct, low-confidence wrong, and medium-confidence (=3).

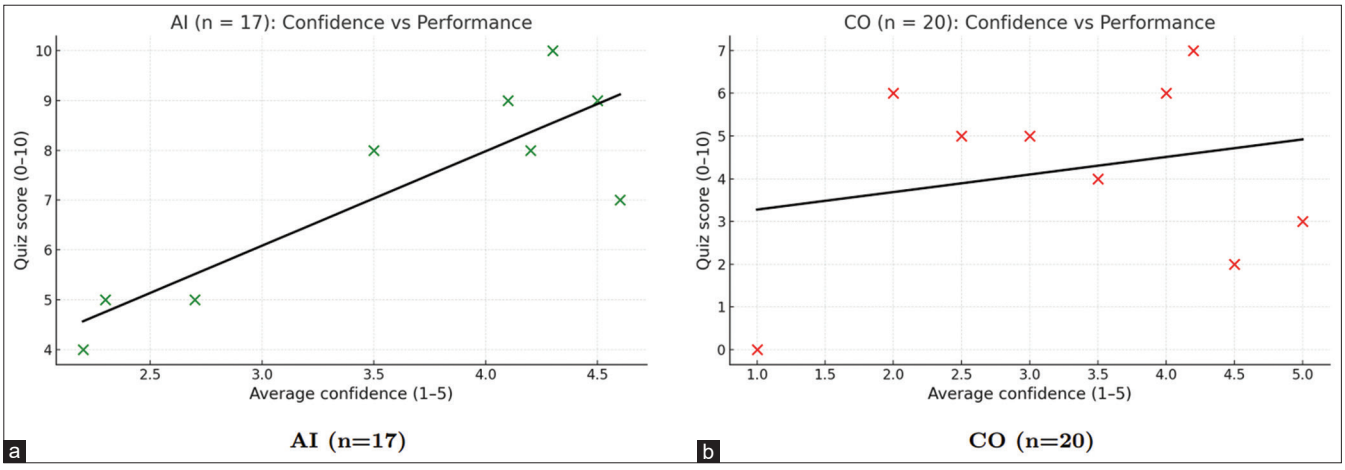


Figure 2: Confidence–accuracy relationship. Each point is one student: x = average confidence across the 10 quiz items (1–5), y = quiz score (0–10). The black line is an OLS fit within each group. (a) AI group and (b) control group

A positive β_1 indicates that students with higher confidence tend to achieve higher scores, while a negative β_1 would indicate the opposite.

Figure 2a (AI; green dots) shows a strong positive relation: Higher average confidence is associated with higher quiz scores (steep positive slope). Figure 2b (CO; red dots) shows a weak/near–zero relation: scores vary widely at each confidence level. Thus, in the CO group, reported confidence does not reliably indicate accuracy.

The AI group exhibits good metacognitive calibration: Students' confidence judgments track their performance, when they feel more confident, they tend to be more correct. The CO group shows poorer calibration: Confidence is largely uninformative (or sometimes misleading) about performance, suggesting overconfidence for some students and underconfidence for others. These results align with the calibration–category analysis (HCC/HCW/LCW/LCC),

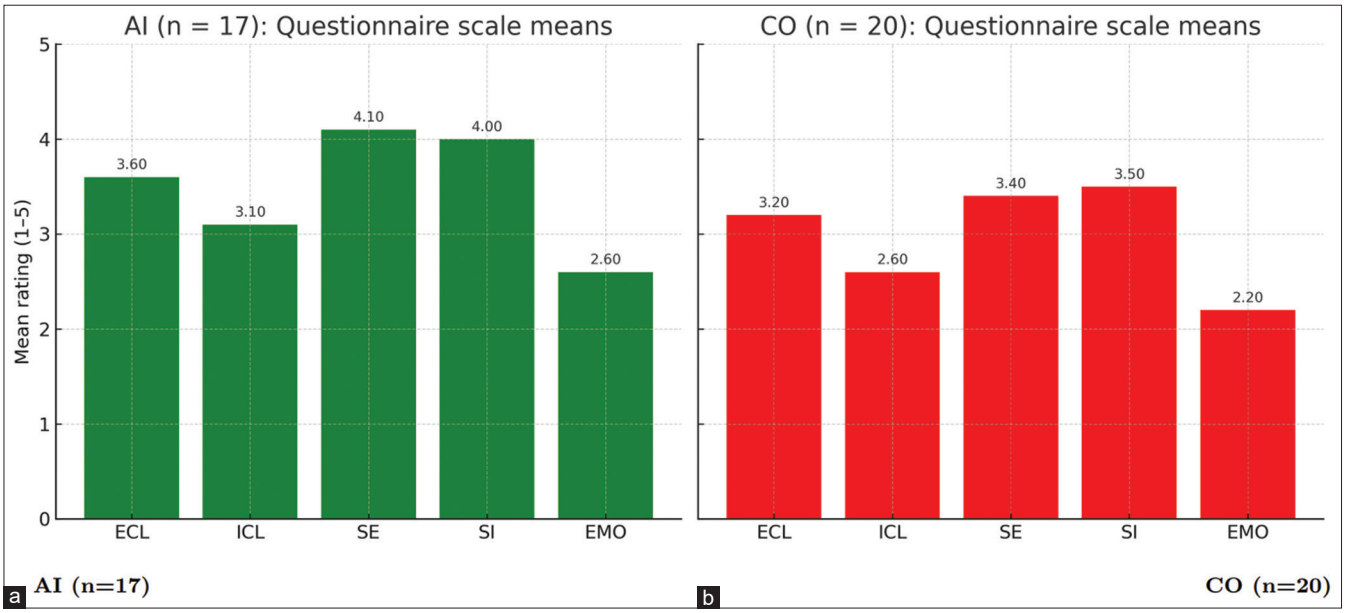
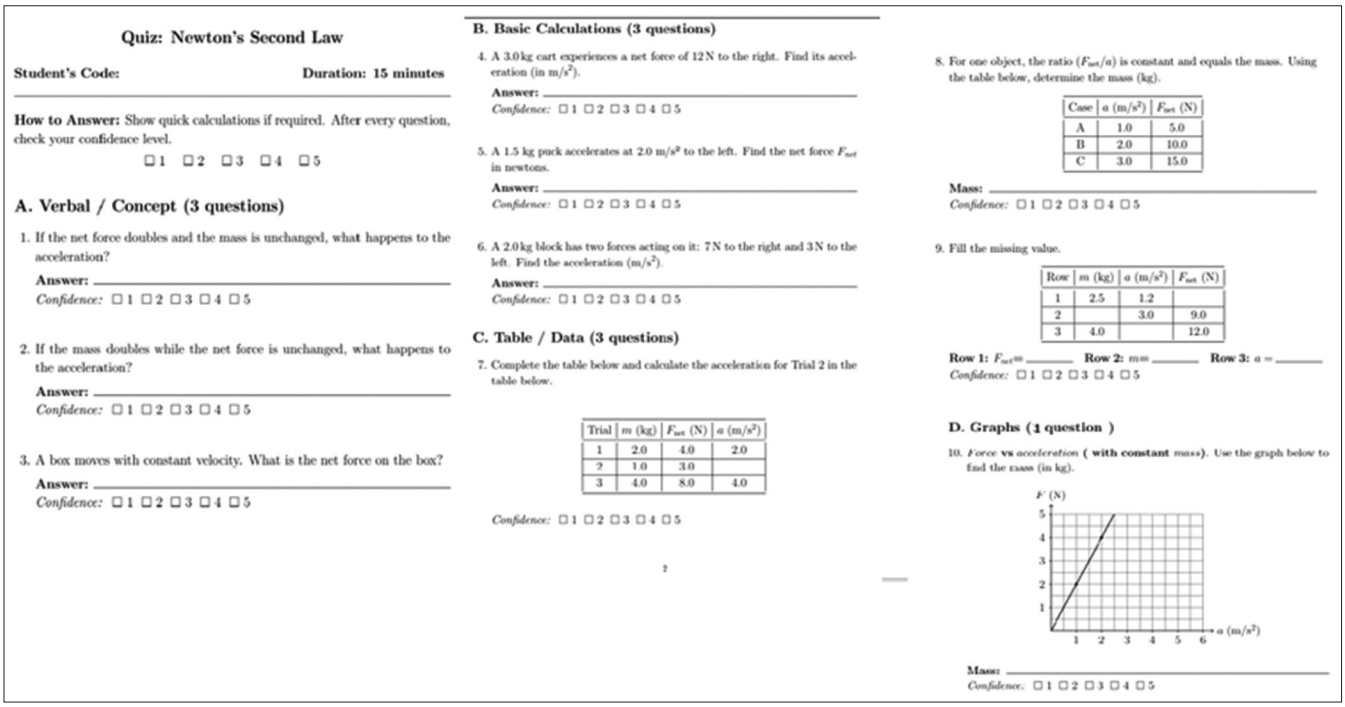


Figure 3: Questionnaire scale means (1–5). (a) AI group and (b) control group. AI group bars are green; control group bars are red; no error bars are shown. ECL: Extraneous cognitive load, ICL: Intrinsic cognitive load, SE: Self-efficacy, SI: Situational interest, EMO: Affect, AI: Artificial intelligence



Quiz: Newton's Second Law

Student's Code: **Duration: 15 minutes**

How to Answer: Show quick calculations if required. After every question, check your confidence level.

☐1 ☐2 ☐3 ☐4 ☐5

A. Verbal / Concept (3 questions)

1. If the net force doubles and the mass is unchanged, what happens to the acceleration?
Answer: _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

2. If the mass doubles while the net force is unchanged, what happens to the acceleration?
Answer: _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

3. A box moves with constant velocity. What is the net force on the box?
Answer: _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

B. Basic Calculations (3 questions)

4. A 3.0 kg cart experiences a net force of 12 N to the right. Find its acceleration (in m/s²).
Answer: _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

5. A 1.5 kg puck accelerates at 2.0 m/s² to the left. Find the net force F_{net} in newtons.
Answer: _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

6. A 2.0 kg block has two forces acting on it: 7 N to the right and 3 N to the left. Find the acceleration (m/s²).
Answer: _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

C. Table / Data (3 questions)

7. Complete the table below and calculate the acceleration for Trial 2 in the table below.

Trial	m (kg)	F_{net} (N)	a (m/s²)
1	2.0	4.0	2.0
2	1.0	3.0	
3	4.0	8.0	4.0

Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

8. For one object, the ratio (F_{net}/a) is constant and equals the mass. Using the table below, determine the mass (kg).

Case	a (m/s²)	F_{net} (N)
A	1.0	5.0
B	2.0	10.0
C	3.0	15.0

Mass: _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

9. Fill the missing value.

Row	m (kg)	a (m/s²)	F_{net} (N)
1	2.5	1.2	
2		3.0	9.0
3	4.0		12.0

Row 1: F_{net} = _____ **Row 2:** m = _____ **Row 3:** a = _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

D. Graphs (1 question)

10. *Force* **vs** *acceleration* **(with constant** *mass***).** Use the graph below to find the mass (in kg).

Mass: _____
Confidence: ☐1 ☐2 ☐3 ☐4 ☐5

Figure 4: Full version of Newton's Second Law quiz used in the study. Each question includes a confidence rating scale to measure students' self-monitoring

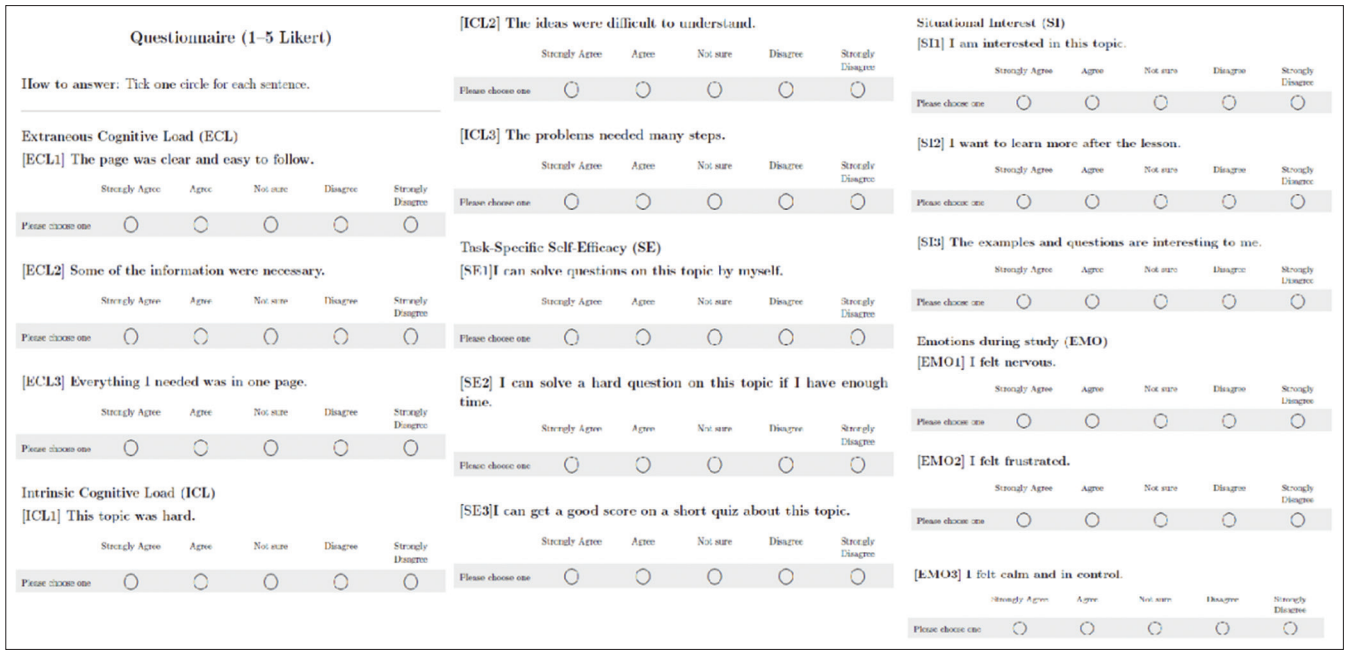


Figure 5: Full post-study questionnaire measuring cognitive load, self-efficacy, situational interest, and emotional responses

Table 2: α for 3-item scales by group

Scale	AI (n=17)	CO (n=20)
ECL	0.546	0.203
ICL	0.641	0.323
SE	0.677	0.342
SI	0.733	0.379
EMO	0.785	0.402

ECL: Extraneous cognitive load, ICL: Intrinsic cognitive load, SE: Self-efficacy, SI: Situational interest, EMO: Affect

indicating that AI tutoring strengthens the link between subjective confidence and objective performance.

Questionnaire and Reliability (Cronbach's α Scales)

After studying (but before the quiz), students completed five brief 3-item scales: ECL, ICL, self-efficacy (SE), SI, and affect (EMO). Items were rated 1–5 (1 = strongly disagree, 5 = strongly agree). For each scale, we averaged the three items per student; Figure 3 reports group means (1–5). Figure 3 shows that with these data, the AI group scores higher than the CO group on all five scales (ECL, ICL, SE, SI, EMO). The largest gaps are in SE and SI, indicating that AI-supported students felt more confident and more interested in the activity. EMO is also higher for the AI group, suggesting a slightly more positive affective experience. ECL and ICL are modestly higher in the AI group as well; this reflects greater perceived processing of the materials and topic demands, yet it coexists with better scores and calibration.

Table 2 illustrates that Cronbach's α for the 3-item scales is consistently higher in the AI group (about 0.55–0.79) than in the control group (about 0.20–0.40). Thus, the AI students' responses hang together better (more internally consistent), whereas the control scales show weak consistency.

CONCLUSION

This study compared a short AI-tutored study session with a matched textbook session for pre-instruction learning on Newton's Second Law. Across all analyses, the AI condition produced better learning and stronger metacognitive monitoring.

AI students achieved higher quiz scores than the control group (6.65/10 vs. 4.45/10). They also showed fewer low-confidence items (L_{AI} = 3.34 vs. L_{CO} = 3.85) and a lower guess rate (G_{AI} = 33.5% vs. 38.5%), indicating less reliance on guessing.

The AI group displayed the desirable calibration profile: Many high-confidence correct responses (HCC ≈ 4.94 of 10) and very few HCW responses (HCW ≈ 0.24). By contrast, the control group showed fewer confidently correct answers (HCC ≈ 2.70) and frequent confidently wrong answers (HCW ≈ 3.30), a hallmark of overconfidence.

At the student level, average confidence strongly predicted quiz score in the AI group (positive regression slope), whereas the relation was weak/near-zero in the control group. Thus, only AI students' expressed confidence reliably tracked their accuracy.

After study (before the quiz), AI students reported higher means on all five scales – Extraneous Load (ECL), Intrinsic Load (ICL), SE, SI, and Affect (EMO) – with the clearest advantages in SE and SI. Internal consistency (Cronbach's α) was consistently higher for AI across scales (~0.55–0.79) than for control (~0.20–0.40), suggesting more coherent, interpretable responses in the AI condition.

A brief AI-guided session improved immediate conceptual performance and, crucially, fostered better metacognitive judgment: students were confident when correct and rarely

confident when wrong. The AI tutor also supported motivation (higher SE and SI) without adding undue perceived difficulty, indicating that conversational AI can simultaneously enhance cognitive and metacognitive outcomes in introductory physics.

REFERENCES

Ambrose, S.A., Bridges, M.W., DiPietro, M., Lovett, M.C., & Norman, M.K. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. San Francisco: Jossey-Bass.

Bewersdorff, A., Zhai, X., Roberts, J., & Nerdel, C. (2023). Myths, mis- and preconceptions of artificial intelligence: A review of the literature. *Computers and Education Artificial Intelligence*, 4, 100143.

Bitzenbauer, P. (2023). ChatGPT in physics education: A pilot study on easy-to-implement activities. *Contemporary Educational Technology*, 15(3), ep430.

Bransford, J.D., Brown, A.L., & Cocking, R.R. (2000). *How People Learn: Brain, Mind, Experience, and School*. Washington, DC: National Academies Press.

Chan, C.K.Y., & Hu, W. (2023). *Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education*. [arXiv Preprint].

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Flavell, J.H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906-911.

Graesser, A.C., Nye, B.D., & Wang, X. (2016). AutoTutor and friends: Learning with conversational agents. *International Journal of Artificial Intelligence in Education*, 26, 124-132.

Griffith, W.T., & Brosing, J. (2023). *Physics of Everyday Phenomena*. 10th ed. United States: McGraw Hill.

Hacker, D.J., Dunlosky, J., & Graesser, A.C. (Eds.). (2009). *Handbook of Metacognition in Education*. United Kingdom: Routledge.

Hake, R.R. (1998). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data. *American Journal of Physics*, 66(1), 64-74.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141-158.

Jia, Z., Zou, Y., Sun, H., Zhang, C., & Tang, J. (2023). ChatGPT and large language models in academia: Advantages, concerns and current trends. *BioData Mining*, 16(1), 42.

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

Mazur, E. (1997). *Peer Instruction: A User's Manual*. Upper Saddle River, NJ: Prentice Hall.

McDermott, L.C., & Shaffer, P.S. (2002). *Tutorials in Introductory Physics*. Upper Saddle River, NJ: Prentice Hall.

Pizzi, G., Scarpi, D., & Pantano, E. (2023). I, chatbot! The impact of anthropomorphism and gaze direction on willingness to disclose personal information. *Psychology and Marketing*, 40(7), 1372-1389.

Schraw, G., & Dennison, R.S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19(4), 460-475.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, 74(1), 107-120.

Sun, X., & Zhou, Y. (2022). *AI Tutors for Introductory Physics Problem Solving*. In: Proceedings of the XXX Conference on STEM Education.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55.

Zimmerman, B.J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64-70.

APPENDIX

Appendix A: Quiz and Questionnaire Instruments

To provide full transparency and allow replication, this appendix presents the complete assessment instruments used in this study. Figure 4 shows the full quiz on Newton's Second Law, administered immediately after the self-study phase. The quiz included conceptual, computational, tabular-data, and graphical interpretation questions, with a confidence rating scale after each item to measure students' metacognitive judgments.

Figure 5 includes the full post-study questionnaire administered before the quiz. It was designed to capture multiple cognitive and affective dimensions of the learning experience, including extraneous cognitive load, intrinsic cognitive load, self-efficacy, situational interest, and emotional state. All items were rated on a five-point Likert scale from strongly disagree to strongly agree.

Appendix B: Computation of Cronbach's α

Cronbach's α was used to measure the internal consistency of each 3-item questionnaire scale. First, students' responses to the three items in each scale were collected using a 1–5 Likert scale. The variance of each individual item was then calculated, followed by the variance of the total score obtained by summing the three items for each student. Finally, Cronbach's α was computed using the standard reliability formula:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_T^2}\right), \tag{B1}$$

Where k = 3 is the number of items, σ_i^2 represents the variance of each item, and σ_T^2 is the variance of the total scale scores.

The coefficient α ranges from 0 to 1, where higher values indicate stronger internal consistency among the questionnaire items.