STUDENT–AI RESEARCH MENTOR COLLABORATION: INITIAL FINDINGS FROM AN INTRODUCTORY RESEARCH METHODS COURSE

Abdullah Saykili
Anadolu University
College of Education Anadolu University, Eskisehir, Türkiye

ABSTRACT

This study reports initial findings on student–AI research mentor collaboration during topic and research question (RQ) development in an undergraduate level research methods course. The focal workflow included two stages: teams first drafted a topic and primary RQ without AI support (Assignment 4.1), then individuals consulted an AI mentor and returned to their teams to reconcile recommendations through an explicit accept/adapt/reject process (Assignment 4.2). Following Assignment 4.2, 58 of 74 students (78.4%) completed a survey including 12 Likert-type items (1–5) and three open-ended prompts. Quantitative results indicated strong perceived learning and refinement value, with the highest endorsement for improved understanding of what constitutes a good research question (Item 10, M = 4.43) and high overall recommendation intentions (Item 11, M = 4.52). Feedback clarity was rated favorably (Item 4, M = 4.19), while perceived project-specificity and comfort disagreeing were comparatively lower and more variable (Items 5–8; lowest for Item 8, M = 3.86). Thematic analysis showed that students primarily used the AI mentor to narrow overly broad topics, increase specificity (participants/context/variables), and refine wording; disagreement episodes centered on generic feedback, over-specification demands, and method-type misclassification, typically resolved through feasibility checks, instructor-aligned criteria, and team deliberation. Findings suggest that pairing an explicit checklist with AI mentoring can support calibrated uptake and learner agency in early research design tasks.

KEYWORDS

Generative AI, Human-AI Collaboration, Generative AI in Education, AIEd, Large Language Models, Mixed Methods Research

1. INTRODUCTION

Generative AI is now widely visible in higher education, and students are increasingly using these tools not only for efficiency but also as interactive partners that resemble tutoring, feedback, and mentoring dialogue (Kasneci et al., 2023; UNESCO, 2023). This creates a pedagogical opportunity: AI can provide rapid "second reader" feedback that may help students notice strengths, detect ambiguities, and consider feasibility. However, large language models can also produce responses that appear confident while being incomplete, biased, or factually wrong (Bender et al., 2021), including fabricated details or sources (hallucinations) (Ji et al., 2023). In this context, the central instructional problem shifts from whether students will use AI to how students can use it responsibly, maintaining critical judgment and learner agency rather than outsourcing decisions. This aligns with calls to move beyond isolated, tool-centric perspectives toward context-sensitive educational research and practice (Zawacki-Richter et al., 2019) and with guidance emphasizing responsible use in education (UNESCO, 2023).

This paper reports initial findings from ARY111 - Introductory Research Methods in Social Sciences undergraduate course. As part of their final research report students first drafted a team topic and RQ without AI and then refined them using an AI research mentor alongside a structured checklist. Survey data were collected from 58 of 74 students (78.4%) using 12 Likert items and three open-ended prompts. The study addresses three questions:

- RQ1: To what extent do students perceive the AI research mentor as supporting learning and refinement of their topic/RQ?

- RQ2: How do students evaluate feedback quality and critical engagement, including clarity, project-specificity, and comfort disagreeing with AI suggestions?
- RQ3: How do students perceive agency, boundary awareness, and acceptability, and what guidance do they offer peers for smart, critical use?

2. LITERATURE REVIEW

Early-stage research design benefits from structured supports that make quality criteria visible and actionable. Formative assessment principles emphasize that effective feedback clarifies standards, supports self-monitoring, and helps learners close the gap between current and desired performance (Nicol & Macfarlane-Dick, 2006). Importantly, students also need "feedback literacy" which refers to the capacity to interpret feedback, judge its relevance, and decide what to do with it rather than simply receiving comments (Carless & Boud, 2018). From a self-regulated learning perspective, novices develop competence through cycles of goal setting, monitoring, and strategy adjustment (Zimmerman, 2002). In practice, checklists and exemplars can externalize expectations and make revision decisions more deliberate, especially when students must justify feasibility and alignment under course constraints.

Recent syntheses note that large language models can support educational work by helping learners generate options, clarify concepts, structure drafts, and revise text (Kasneci et al., 2023; UNESCO, 2023). Yet educational value depends heavily on how AI use is framed, verified, and integrated into disciplinary standards (UNESCO, 2023). Risks are non-trivial: language model outputs can be persuasive even when they embed problematic assumptions or errors (Bender et al., 2021), and hallucination remains a known failure mode in natural language generation (Ji et al., 2023). In parallel, academic integrity discussions increasingly argue for proactive instructional redesign guiding responsible use and making learning processes visible rather than relying only on punitive or detection-oriented responses (Cotton et al., 2024). For early adoption in courses, students' perceptions and practical use-strategies therefore become central evidence, not a side note.

A key question in human–AI collaboration is appropriate reliance: when users trust and use automated suggestions, when they calibrate, and when they resist (Lee & See, 2004; Parasuraman & Riley, 1997). Guidance for human–AI interaction similarly emphasizes that systems should support user control and informed decision-making (Amershi et al., 2019). Applied to learning, this implies that AI should function as a prompt for reasoning rather than a substitute for judgment, particularly when tasks require feasibility decisions and ethical constraints. In an educational setting, this perspective foregrounds whether students can (a) evaluate specificity (e.g., avoiding generic feedback), (b) disagree with AI recommendations when warranted, and (c) remain the main decision-maker.

Despite growing discussion of generative AI in education, research still needs more context-sensitive, educator- and learner-centered accounts of how students collaborate with AI across authentic, sequential course workflows (Zawacki-Richter et al., 2019; UNESCO, 2023). This study contributes initial evidence from a checklist-guided, team-based topic/RQ refinement process, focusing on student-reported mechanisms of use, critique, and decision-making.

3. METHOD

3.1 Research Design

This study used a cross-sectional mixed-methods survey design to document students' initial experiences collaborating with an AI research mentor during topic and research question (RQ) refinement. The quantitative strand comprised descriptive ratings on Likert-type items assessing perceived instructional value, feedback quality/critical engagement, and agency/boundary awareness. The qualitative strand comprised open-ended responses eliciting concrete help mechanisms, disagreement episodes and decision rules, and peer-facing advice. The two strands were integrated at the interpretation stage, enabling convergence and complementarity between patterns in item ratings and thematic evidence (Creswell & Plano Clark, 2017).

3.2 Context and Assignment Design

The study was conducted in ARY111: Research Methods in Social Sciences, an undergraduate introductory course in which students develop and report a small-scale research project. The focal instructional sequence addressed in this study was Assignment 4, structured as a two-stage workflow to distinguish team-based work from subsequent AI-supported refinement.

In Assignment 4.1, students worked as a team without GenAI. Teams met to brainstorm possible topics, narrow a broad interest area into a feasible topic statement, draft a primary research question (and optional sub-questions), and evaluate the topic–RQ pairing using a Researchable Topic & Good Research Question Checklist. Teams documented attendance, alternatives considered, decisions, and rationales in a meeting memo.

In Assignment 4.2, students engaged in a structured human–AI collaboration sequence. Each student individually used the AI mentor to review the team's topic/RQ, compared team checklist ratings with the AI mentor's ratings in an individual comparison table, and then returned to the team for a second meeting. Teams discussed AI recommendations, decided what to accept, adapt, or reject (with reasons), and finalized the topic and primary RQ, documenting the process in a Task 2 meeting memo. This design foregrounds calibrated uptake of AI feedback under constraints, rather than passive receipt (Amershi et al., 2019; Lee & See, 2004).

3.3 Participants and Data Source

The course enrolled 74 undergraduate students. Following completion of Assignment 4.2, students were invited to complete an anonymous, voluntary survey about their experience using the AI research mentor for topic/RQ refinement. The survey yielded 58 responses (78.4% response rate). The dataset comprises responses to 12 Likert-type items and three open-ended prompts. Analyses were conducted using available responses per item.

3.4 Measures

All Likert-type items were rated on a 5-point scale (1 = Strongly disagree; 5 = Strongly agree). For alignment with the research questions and parsimonious reporting, items were grouped into three analytic clusters.

The RQ1 cluster (learning/refinement value) comprised Items 1, 2, 3, 6, and 10, capturing perceived diagnostic support (noticing strengths; identifying weaknesses/unclear elements), reflective value of comparing checklist ratings with AI ratings, confidence when using the checklist with the AI mentor, and overall perceived improvement in understanding what constitutes a good research question. Conceptually, this cluster aligns with feedback functions that clarify standards and support self-monitoring and revision decisions (Carless & Boud, 2018; Nicol & Macfarlane-Dick, 2006).

The RQ2 cluster (feedback quality and critical engagement) comprised Items 4, 5, 7, and 8, assessing perceived clarity of the AI mentor's feedback, perceived project-specificity (vs generic advice), perceived encouragement of critical thinking rather than answer-giving, and comfort disagreeing with AI suggestions when justified, which are dimensions linked to appropriate reliance and calibration in human–automation interaction (Lee & See, 2004; Parasuraman & Riley, 1997).

The RQ3 cluster (agency/limits/acceptability) comprised Items 9, 11, and 12, assessing perceived respect for ethical/practical constraints, willingness to recommend AI mentoring to peers, and perceived ability to remain the primary decision-maker when using AI tools, which are considered central to responsible human–AI interaction in education (Amershi et al., 2019; UNESCO, 2023).

Open-ended prompts asked students to report (Q1) the most helpful specific contribution of the AI mentor, (Q2) disagreements or unhelpful/unrealistic feedback and decision processes, and (Q3) advice for future students on smart, responsible, and critical use.

3.5 Data Analysis

Quantitative analyses summarized item responses using means and standard deviations, alongside endorsement rates (percentage selecting 4–5). When useful for concise reporting, cluster-level summaries were used descriptively to reflect RQ-aligned patterns. Qualitative responses were analyzed using a thematic content

analytic approach organized initially by prompt focus (help mechanisms; disagreement and decision rules; peer advice) and iteratively refined to capture recurrent categories across responses. Mixed-method integration was operationalized through a compact joint display that juxtaposed the most prominent quantitative signals (e.g., highest/lowest endorsed items within clusters) with corresponding qualitative themes and illustrative excerpts, supporting interpretive triangulation (Creswell & Plano Clark, 2017).

4. RESULTS

4.1 RQ1: Learning and Refinement Value

Across the learning/refinement cluster, students reported consistently high perceived value (RQ1 composite M = 4.20, SD = 0.56) (See Table 1). The highest item in this cluster was improved understanding of what a good research question looks like (Item 10, M = 4.43, SD = 0.73; median = 5), followed by perceived gains in confidence when using the checklist together with the AI mentor (Item 6, M = 4.21) and noticing strengths in their draft (Item 1, M = 4.21). Item distributions were generally negatively skewed, indicating responses concentrated toward agreement rather than neutrality or disagreement.

Qualitative responses (Open-Ended Q1) clarified how students experienced this value. The modal mechanism was scope reduction recognizing that an initial topic/RQ was too broad and narrowing it to a researchable slice. For instance, a student stated: *"It made us realize that our topic was very broad, and we narrowed it down immediately."* A second mechanism was increasing specificity and measurability (participants, context, variables), often framed as converting vague phrasing into explicit operational terms (e.g., specifying platform/context and clarifying variables), which can be observed in a student statement: *"The AI mentor made the context and the participants more specific."* Students also emphasized iterative rewriting for precision, describing repeated revision cycles to sharpen wording and focus: *"Using AI taught me what a good research question should be… how specific it should be… We edited our sentence and question again and again."* Collectively, the qualitative evidence suggests that the AI mentor was experienced primarily as a diagnostic and refinement scaffold that supported narrowing, operationalization, and clarity. This finding is consistent with the high mean levels observed for Items 1–3, 6, and 10.

4.2 RQ2: Feedback Quality and Critical Engagement

Perceived feedback quality and critical engagement were positive but comparatively lower than RQ1 and RQ3 (RQ2 composite M = 3.99, SD = 0.62). Students rated clarity of the AI mentor's feedback highly (Item 4, M = 4.19, SD = 0.96; median = 4.5). In contrast, perceived project-specificity was lower (Item 5, M = 3.95), and the lowest mean in the entire instrument was comfort disagreeing with AI suggestions (Item 8, M = 3.86, SD = 1.03), indicating greater dispersion in students' experiences of disagreement and contestation.

Open-Ended Q2 explains these differences. While many students reported no disagreement *("We accepted it all")*, disagreement clustered around three recurring issues. First, students described feedback as too general or insufficiently diagnostic, leading to selective uptake: *"Sometimes the suggestions were too general… I compared them with my project requirements and chose the parts that were most relevant."* Second, some experienced over-specification demands (e.g., pushing narrower platforms or extra parameters), which they evaluated as unnecessary or impractical; these cases were typically resolved via team discussion and adaptation rather than wholesale acceptance. Third, students noted misalignment or misinterpretation, where the AI flagged missing details that were already present, prompting them to retain original wording: *"We had already included this detail… the feedback didn't fully apply, so I decided to keep my original wording."*

Across disagreement cases, students articulated consistent decision rules: alignment with assignment instructions/instructor guidance, consultation with teammates, feasibility checks (time/recruitment/access), and an "accept–adapt–reject" logic. These reported adjudication practices align with the instrument's focus on critical engagement (Item 7, M = 3.95) and help interpret why disagreement comfort (Item 8) showed the greatest variability.

Table 1. Descriptive statistics for survey items

Cluster	Item	M	SD	Median
RQ1 Learning/refinement	1. The AI mentor helped me notice strengths in our research topic and research question.	4.21	0.72	4.00
	2. The AI mentor helped me identify weaknesses or unclear parts in our topic and/or research question.	4.07	1.02	4.00
	3. Comparing my own checklist ratings with the AI mentor's ratings helped me think more carefully about what makes a topic "researchable."	4.07	0.81	4.00
	6. Using the checklist (A1–B3) together with the AI mentor made me feel more confident about our topic and research question.	4.21	0.83	4.00
	10. Overall, using the AI mentor improved my understanding of what a good research question looks like.	4.43	0.73	5.00
	RQ1 Composite	*4.20*	*0.56*	*4.20*
RQ2 Feedback/critical engagement	4. The AI mentor's feedback was clear and easy for me to understand.	4.19	0.96	4.50
	5. The AI mentor's feedback felt specific to our project (not just generic advice).	3.95	0.96	4.00
	7. The AI mentor encouraged me to think critically instead of just giving me the final answer.	3.95	0.93	4.00
	8. I felt comfortable disagreeing with the AI mentor's ratings or suggestions when I had good reasons.	3.86	1.03	4.00
	RQ2 composite	*3.99*	*0.62*	*4.00*
RQ3 Agency/limits/recommendation	9. The AI mentor respected ethical and practical limits (for example, realistic sample size, access to participants, time limits).	4.14	0.95	4.00
	11. Overall, I would recommend using an AI mentor for research projects to other first-year students.	4.52	0.84	5.00
	12. After this experience, I believe I can remain the main decision-maker in my research, even when I use AI tools.	4.33	0.76	4.50
	RQ3 composite	*4.33*	*0.61*	*4.33*

Observed cluster-level endorsement (composite ≥ 4.0): RQ1 = 63.8%; RQ2 = 58.6%; RQ3 = 82.8%.

4.3 RQ3: Agency, Limits and Recommendation

Students reported strong outcomes for agency/limits/recommendation (RQ3 composite M = 4.33, SD = 0.61), with particularly high endorsement of recommending an AI mentor to other first-year students (Item 11, M = 4.52, SD = 0.84; median = 5). Perceived ability to remain the main decision-maker was also high (Item 12, M = 4.33, SD = 0.76), and perceived respect for ethical/practical limits was favorable (Item 9, M = 4.14). Notably, cluster distributions showed the largest concentration at the upper end for RQ3 (82.8% of composite scores ≥ 4.0).

Open-Ended Q3 adds detail on what "smart and critical use" meant to students. The dominant advice was to treat AI as a support tool rather than an authority, preserving human control: *"Use AI as an assistant, not as a replacement for your own thinking."* A second theme emphasized verification and triangulation *("Always double check the answers… compare them with sources")*, reflecting awareness of reliability limits. A third theme focused on prompting tactics (clear, specific prompts; follow-up questions; requesting alternatives) as a prerequisite for useful feedback. Finally, a subset offered sophisticated workflow guidance: draft first, then use AI as a structured checker for clarity, feasibility, and ethics using criteria such as the checklist while explicitly warning against overreliance and "laziness." Together, these themes triangulate the high quantitative endorsement of recommendation (Item 11), agency (Item 12), and limits (Item 9), while also clarifying the normative practices students viewed as necessary for responsible adoption.

Table 2. Joint display for mixed-method integration

RQ Cluster	Prominent Quan. Signals	Convergent Qual. Themes	Illustrative Excerpts
RQ1: Learning & refinement value (Items 1,2,3,6,10)	Highest: Item 10 "good RQ understanding" (M = 4.43, SD = 0.73). High: Items 1 and 6 (M = 4.21 each). Lowest within cluster: Items 2 and 3 (M = 4.07 each).	Scope reduction (recognizing over-breadth → narrowing). RQ precision & rewriting (clearer wording, iterative revision). Operationalization (making participants/context/variables explicit).	"It made us realize that our topic was very broad, and we narrowed it down immediately." / "Using AI taught me what a good research question should be… We edited our sentence and question again and again."
RQ2: Feedback quality & critical engagement (Items 4,5,7,8)	Highest: Item 4 "feedback clarity" (M = 4.19, SD = 0.96). Lower: Items 5 and 7 (M = 3.95). Lowest: Item 8 "comfortable disagreeing" (M = 3.86, SD = 1.03; greatest dispersion).	Generic/overly positive feedback reduced usefulness; students used selective uptake. Over-specification requests (platform/time/sample) sometimes judged unnecessary. Misfit/misinterpretation prompted rejection/retention of original wording. Method-type ambiguity (quant vs qual) triggered debate.	"Sometimes the suggestions were too general… I chose the parts that were most relevant." / "We had already included this detail… the feedback didn't fully apply, so I decided to keep my original wording."
RQ3: Agency, limits & recommendation (Items 9,11,12)	Highest: Item 11 "recommend to other first-years" (M = 4.52, SD = 0.84). High: Item 12 "remain main decision-maker" (M = 4.33, SD = 0.76). Favorable: Item 9 "respected limits" (M = 4.14, SD = 0.95).	Advice centered on agency preservation (AI as assistant, not authority), verification/triangulation (double-checking), and prompt discipline (specific prompts, follow-ups, alternatives). Students recommended using AI as a structured checker for feasibility/ethics, not as final arbiter.	"Use AI as an assistant, not as a replacement for your own thinking." / "Always double check the answers… compare them with sources and think critically."

Quantitative patterns show strong perceived learning value and high recommendation/agency, while qualitative data clarify *how* value was produced (narrowing, operationalizing, rewriting) and *where* friction occurred (generic feedback, over-specification, misfit, and method-type ambiguity) (see Table 2). The decision process students described—check requirements → feasibility-check → team deliberation → accept/adapt/reject—helps explain why "comfort disagreeing" was the most variable item within the feedback/critical-engagement cluster.

5. DISCUSSION

Across RQ1, students' ratings and examples converge on the AI mentor functioning as a rapid second reader for early-stage research design. Students most often described how AI support helped them diagnose over-breadth, strengthen specificity, and iteratively rewrite questions into more researchable forms. This pattern is consistent with feedback models that emphasize clarifying standards and enabling learners to close gaps through iterative revision (Nicol & Macfarlane-Dick, 2006) and with the notion of feedback literacy, where students actively interpret and act on feedback rather than simply receive it (Carless & Boud, 2018). In our data, the checklist appeared to make quality criteria explicit, while the AI mentor supplied an additional perspective that prompted refinement decisions.

For RQ2, students rated the AI mentor's feedback as clear but were more mixed about project-specificity and about disagreeing with AI suggestions. Qualitative accounts explain this dispersion. Some students experienced generic or overly positive feedback, others perceived "over-specification" demands, and a subset identified misalignment with assignment requirements. These patterns align with broader cautions that large language model outputs can be persuasive yet incomplete or incorrect, making verification essential (Bender et al., 2021; Ji et al., 2023). From a human–automation perspective, students' reports of selectively accepting, adapting, or rejecting AI suggestions reflects attempts at calibrated reliance (Lee & See, 2004; Parasuraman & Riley, 1997). The strongest decision heuristics were alignment with course instructions, feasibility checks, and team consultation suggesting that the social and procedural scaffolds around AI use matter as much as the tool itself.

RQ3 findings indicate that students largely endorsed agency and recommendation, and their advice emphasized verification, clear prompting, and preserving human control. This is congruent with guidance that responsible generative AI use should be treated as a contextual practice supported by instructional design and guardrails, not merely individual ethics statements (UNESCO, 2023), and with calls for educator- and learner-centered studies situated in authentic learning contexts (Zawacki-Richter et al., 2019). Practically, the results support a replicable pattern for introductory methods courses: draft first → checklist-based self/peer review → AI mentor as second opinion → team reconciliation → documented accept/adapt/reject decisions, which operationalizes "AI as support" rather than "AI as authority" (Amershi et al., 2019).

6. CONCLUSION

This paper provides initial, classroom-based evidence on how undergraduates collaborate with an AI "research mentor" during topic and research question refinement. The combined quantitative and qualitative findings suggest that the checklist + AI pairing can promote a reflection loop in which students diagnose over-breadth, increase specificity, and revise toward feasibility while maintaining decision authority. The observed emphasis on verification and selective uptake is notable given established concerns about confident but unreliable outputs and hallucinated content (Bender et al., 2021; Ji et al., 2023) and reinforces the importance of designing for appropriate reliance rather than unconditional trust (Lee & See, 2004; Parasuraman & Riley, 1997). More broadly, the study responds to calls for context-sensitive research that examines how AI is used in real educational workflows rather than in isolated, tool-centric evaluations (Zawacki-Richter et al., 2019), and it aligns with policy-oriented guidance emphasizing proactive instructional strategies for responsible use (UNESCO, 2023).

Several limitations constrain interpretation. Data come from a single course and rely on self-report perceptions; the study did not include an independent, rubric-based rating of topic/RQ quality or learning performance outcomes. In addition, the design is inherently sequential: students completed the team-only draft before the AI-supported refinement, so learning progression and practice effects cannot be separated from AI exposure. Future work should incorporate objective quality scoring of drafts (pre/post), link perceptions to performance indicators (e.g., rubric ratings or downstream project quality), and test alternative designs (e.g., cohort comparisons or counterbalancing). Even with these constraints, the study offers a practical, defensible workflow for introductory research methods: pairing a checklist with AI mentoring and requiring teams to document accept/adapt/reject decisions, thereby turning AI use into an accountable learning practice.

ACKNOWLEDGEMENT

This work has been supported by Anadolu University Scientific Research Projects Coordination Unit under grant number #SBA-2026-3302.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human–AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300233

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2190148

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*. https://doi.org/10.1145/3571730

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., & Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253. https://doi.org/10.1518/001872097778543886

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. https://doi.org/10.1186/s41239-019-0171-0

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice, 41*(2), 64–70. https://doi.org/10.1207/S15430421TIP4102_2