Journal of Innovative Learning
Institute for Innovative Learning, Mahidol University
Volume 2, Issue 2, page 28-42

Technical Article

Designing AI-Assisted Learning for Evidence-Grounded Academic Self-Efficacy: A Technical Framework for Higher Education

Chanakan SOJAYAPAN*
Department of Educational Technology, Faculty of Education, Silpakorn University, Thailand
* Corresponding author: chanakan.soj@gmail.com

Abstract: This technical article proposes a learning design and assessment framework that separates AI-assisted performance, independent performance, and students' beliefs about their academic capability. A purposive narrative synthesis connects self-efficacy theory with selected higher education studies on capability beliefs, feedback use, and task performance. The conceptual framework specifies four elements: the target capability, permitted AI assistance, evidence of learner decisions, and separately measured belief and performance outcomes. Its five-phase implementation sequence translates these elements into activities: specify, attempt, evaluate, revise, and demonstrate. A worked academic writing example includes repeated capability ratings for the same task and assistance condition, a separate rating for a later task without AI, and provisional writing criteria. The article also explains how the order and withdrawal of AI assistance affect the interpretation of assessment evidence. The framework is a design proposal, not an empirically validated intervention. Future studies should test assistance sequences, the correspondence between capability judgments and performance, and feasibility across disciplines.

Keywords: Generative AI, academic self-efficacy, learning design, AI-assisted learning, higher education

Introduction

Generative artificial intelligence (Generative AI) has changed how students in higher education access academic support. Students can use it to seek explanations, obtain feedback on their work, and generate examples. Generative AI thus serves not only as a source of information but also as a tool that can support students' thinking and task performance (Miao & Holmes, 2023). At the same time, its capacity to help students complete tasks raises a question: To what extent does success achieved with AI assistance reflect students' own capabilities, and how do students distinguish their contributions from those of the technology?

This question is particularly relevant to self-efficacy, which refers to individuals' beliefs in their capabilities to organize and carry out the actions needed to attain particular outcomes (Bandura, 1997). Self-efficacy is more specific than a general sense of confidence. It concerns judgments of capability in relation to particular tasks and contexts and can influence students' choices, effort, and persistence when they meet difficulty (Bandura, 1997; Usher et al., 2023).

When Generative AI contributes to task completion, the resulting work may reflect students' knowledge and strategies alongside the assistance provided by the technology. Studies in writing and programming have reported improvements in self-efficacy specific to those domains (Huang & Mizumoto, 2024; Yilmaz & Karaoglan Yilmaz, 2023). However, a comparison of teacher and Generative AI assistance found that students assisted by AI achieved higher learning outcomes, while those assisted by a teacher reported higher self-efficacy (Yang et al., 2026). These findings raise two related questions: What evidence

do students draw on when judging their capabilities, and what does success achieved with assistance reveal about their learning?

This article synthesizes scholarship on academic self-efficacy and examines how Generative AI may shape students' judgments of their capabilities in higher education. It considers the challenges of interpreting assisted task success alongside students' perceived and demonstrated capabilities, then proposes directions for learning design and future research. Average effects of Generative AI on university students' learning outcomes have been synthesized meta-analytically (Chen & Cheung, 2025). This article instead focuses on what current evidence can and cannot tell instructors about students' capabilities, and on how learning and assessment can be designed around that evidence.

Literature Review

Self-Efficacy and Learning

Definition and Significance

Self-efficacy is a central concept in Bandura's social cognitive theory. It refers to individuals' beliefs in their capabilities to organize and carry out the actions required to attain particular outcomes. These beliefs are specific to domains and tasks: a learner may feel capable of completing one academic task but uncertain about another. Measures of self-efficacy should describe the relevant task and its level of challenge, rather than ask only about confidence in general (Bandura, 1997, 2006).

In higher education, academic self-efficacy concerns students' beliefs about their capability to perform specified academic tasks, such as analyzing data, constructing an argument, or solving a problem in their discipline. Although related to motivation and learning behavior, it is distinct from both. Self-efficacy may shape the tasks students choose and the effort they sustain, while their performance and subsequent interpretation of that performance may, in turn, inform their judgments of capability (Honicke et al., 2023; Zimmerman, 2000).

The study of Generative AI also requires a distinction between academic self-efficacy and AI self-efficacy. Y.-Y. Wang and Chuang (2024) identified four factors in an AI self-efficacy scale: using AI for assistance, anthropomorphic interaction, comfort with AI, and technological skills. Confidence in using an AI system does not necessarily imply equal confidence in analyzing evidence, writing, or solving an academic problem. Researchers should specify whether a measure concerns students' capability to perform an academic task, use AI, or complete the task with AI assistance. Without this distinction, changes in self-efficacy may be difficult to interpret (Y.-Y. Wang & Chuang, 2024).

Development of Self-Efficacy

Bandura (1997) identified four principal sources of self-efficacy information: mastery experiences, vicarious experiences, social persuasion, and physiological and affective states. Their influence depends on how individuals interpret the information these experiences provide. In educational settings, instructors can support this interpretive process by showing students their progress, providing appropriate feedback, offering opportunities to learn from others, and helping students respond to the emotional demands of learning (Usher et al., 2023).

Mastery experiences are particularly relevant because they provide information about what learners have accomplished. Their implications for self-efficacy, however, depend partly on how learners explain that accomplishment. Success attributed to one's own knowledge, strategies, and decisions may be interpreted differently from success attributed primarily to external assistance. Applied to Generative AI, this theoretical distinction calls

for attention to how students describe both their own contributions and the assistance they received. The quality of a completed product alone cannot establish how students interpreted the experience or what they believe they can do independently (Bandura, 1997; Usher et al., 2023).

Generative AI and Academic Self-Efficacy

Mastery Experiences

Studies of specific learning tasks provide initial evidence that Generative AI may support domain-specific self-efficacy. Yilmaz and Karaoglan Yilmaz (2023) conducted a five-week experiment with 45 undergraduate students in a programming course. Students in the experimental group used ChatGPT during weekly programming practice, whereas those in the control group did not. The ChatGPT group reported higher programming self-efficacy and motivation and obtained higher computational-thinking scores than the control group. The study links AI-assisted practice to outcomes within programming education, although its findings concern a particular course, task, and group of students.

Huang and Mizumoto (2024) examined a different learning context involving 327 Japanese university students. They compared peer feedback with ChatGPT feedback delivered through predefined prompts and found an improvement in writing self-efficacy among students who used ChatGPT. Considered alongside the programming study, this finding suggests that the form of AI assistance and the academic task both matter when interpreting self-efficacy outcomes. The studies measured students' beliefs about their capabilities in writing or programming. Whether an increase extends to academic tasks across disciplines, or to performance without AI assistance, is not yet known. For the present article, they show why it is worth asking how assisted practice shapes students' judgments of their capability and how those judgments relate to what students can do independently.

Feedback and Learning Support

Generative AI can provide feedback while students are developing their work, allowing them to identify possible errors, consider alternative approaches, and revise before submitting a final product. Feedback supports understanding when students judge whether a suggestion addresses the task, is supported by relevant evidence, and improves the quality of their work. Bearman et al. (2024) frame this capacity as evaluative judgement and argue that, in the context of Generative AI, students need to evaluate both AI-generated outputs and the processes through which those outputs are produced. Feedback activities should give students opportunities to compare suggestions with task criteria and explain why they accept, adapt, or reject them.

How students seek and process assistance is also relevant. In a study of 38 students revising essays, A. Chen et al. (2025) found that the ChatGPT group followed a more nonlinear help-seeking process and sometimes skipped feedback evaluation. By comparison, students receiving help from a human expert were more likely to evaluate and process the feedback they received. The study describes differences in the revision process without measuring self-efficacy. These differences suggest a reason to make the evaluation of AI feedback an explicit part of a learning activity.

In practice, an instructor might ask students to identify the problem they want feedback on, check an AI suggestion against disciplinary criteria or source material, and explain the decision reflected in their revision. These proposed steps, which A. Chen et al. (2025) did not test, keep students responsible for judging and using the assistance. This matters for self-efficacy research because a student's belief about what they can do may not correspond to the quality of work completed with help. Future studies should examine

students' feedback decisions, self-efficacy, and performance as separate outcomes (A. Chen et al., 2025; Yang et al., 2026).

Comparison of Evidence

Table 1 Empirical studies of AI-supported learning: Design, findings, and limits

Study	Context and design	Findings used	Reason for inclusion and limits
Yilmaz and Karaoglan Yilmaz (2023)	45 undergraduates; programming experiment comparing ChatGPT use with a control condition	Positive effect on programming self-efficacy	Direct self-efficacy evidence in programming; limited to a specific course and skill.
Huang and Mizumoto (2024)	327 Japanese university students; peer feedback compared with ChatGPT feedback using predefined prompts	Improvement in writing self-efficacy	Direct self-efficacy evidence in writing with guided feedback; limited to one language-learning context.
Yang et al. (2026)	20 students participated in teacher-assisted and AI-assisted conditions	Teacher assistance yielded higher self-efficacy; AI assistance yielded higher learning outcomes	Direct comparison of self-efficacy and learning outcomes; small sample limits generalization.
Y. Wang (2026)	Survey of 744 university students; mediation and moderation analyses	Associations consistent with the study's proposed model	Associational evidence linking Generative AI literacy and academic self-efficacy; cannot establish causality.
Zhang et al. (2024)	Cross-sectional survey of 300 university students	Relationship between academic self-efficacy and AI dependency involved academic stress and performance expectations in the examined model	Self-efficacy measured, but only in relation to AI dependency; cross-sectional data leave the direction of the relationship unclear.
Fan et al. (2025)	Randomized writing-task comparison involving 117 students and four support conditions	The ChatGPT group showed greater essay-score improvement, without significant differences in knowledge gain or transfer	Performance and transfer evidence supporting separate assessments; no academic self-efficacy outcome.
A. Chen et al. (2025)	38 students revised essays with ChatGPT or human-expert assistance	The ChatGPT group more often skipped feedback evaluation	Process evidence motivating explicit feedback evaluation; no self-efficacy outcome.
Dai (2026)	Four-condition writing study with 120 undergraduates at one university; feedback-literacy scripting, calibration training, and a combined condition	Combined condition produced the strongest writing gains and retention after AI support was removed; calibration training was more closely associated with self-assessment accuracy	Writing performance and self-assessment accuracy after support removal; no academic self-efficacy outcome; one university and a baseline imbalance in self-assessment error.

Study	Context and design	Findings used	Reason for inclusion and limits
X. Chen (2026)	Eight-week study in six intact classes; bounded AI assistance with required reflection compared with open AI collaboration	Bounded-assistance group scored higher on an independent task without AI	Independent no-AI performance evidence; no self-efficacy outcome; intact classes and combined supports prevent isolating each design element.

Note. The table lists the primary empirical studies of AI-supported learning cited in this article; it was not intended to represent all available research. The first three studies measured self-efficacy in specific learning contexts; Y. Wang (2026) examined academic self-efficacy within a survey model, and Zhang et al. (2024) examined it in relation to AI dependency. Fan et al. (2025), A. Chen et al. (2025), Dai (2026), and X. Chen (2026) inform design decisions but do not provide direct evidence of academic self-efficacy change.

The studies inform different parts of the proposed framework, although none tests the framework in its entirety. The programming and writing studies support examining capability beliefs in relation to defined tasks and forms of assistance (Huang & Mizumoto, 2024; Yilmaz & Karaoglan Yilmaz, 2023). The teacher–AI comparison illustrates why perceived capability and learning outcomes should be measured separately (Yang et al., 2026). Findings on writing performance and help-seeking provide grounds for including a transfer task and an explicit feedback-evaluation step in future studies (A. Chen et al., 2025; Fan et al., 2025). The survey findings identify an association worth investigating further (Y. Wang, 2026).

Two more recent writing studies sharpen the distinction between performance with AI support and performance after that support is removed. In a four-condition study of 120 undergraduates, feedback-literacy scripting was more closely associated with writing quality and deeper revision, whereas calibration training was more closely associated with self-assessment accuracy. The combined condition produced the strongest writing gains and retention after AI support was removed, although it did not outperform calibration training alone on self-assessment accuracy (Dai, 2026). The study was conducted at one university and reported a baseline imbalance in self-assessment error. In a separate eight-week study, students receiving bounded AI assistance with required reflection scored higher on an independent task without AI than those receiving open AI collaboration. Because the study used six intact classes and combined several forms of support, the contribution of each design element cannot be isolated (X. Chen, 2026). Neither study measured academic self-efficacy; their findings support collecting evidence of students' feedback decisions and testing performance under clearly stated assistance conditions.

Challenges in the Generative AI Era

Task Success Versus Learner Capability

A central challenge is determining what successful completion of an AI-assisted task reveals about a student's capability. Generative AI can propose answers, identify possible errors, and suggest revisions. The quality of the resulting work may reflect several contributions that cannot be separated by examining the final product alone. Fan et al. (2025) illustrate this concern: although the ChatGPT group showed greater improvement in essay scores, differences in knowledge gain and transfer were not statistically significant. Assessments intended to capture learning should consider students' reasoning, decisions, and ability to explain or apply what they have produced, alongside the quality of the product (Bearman et al., 2024; Fan et al., 2025).

Calibrating Self-Efficacy to Capability

Calibration refers to how closely students' judgments of their capabilities correspond to what they can demonstrate on a specified task. Because self-efficacy is specific to tasks and contexts, these judgments should be assessed against clearly defined performance criteria and assistance conditions (Bandura, 1997, 2006). In Generative AI contexts, confidence in completing a task with AI support should be distinguished from confidence in performing it independently. A positive correlation between self-efficacy and performance does not necessarily indicate accurate calibration, as students may still systematically overestimate or underestimate their capabilities.

Empirical findings reinforce the importance of examining these outcomes separately. Yang et al. (2026) measured learning outcomes and self-efficacy separately but did not directly assess calibration. Dai (2026) likewise distinguished writing improvement from self-assessment accuracy, which is related to, but distinct from, self-efficacy calibration. These findings support examining perceived capability and demonstrated performance separately. The present article proposes comparing students' judgments with performance evidence under specified assistance conditions to investigate how accurately those judgments reflect their capabilities.

Dependence on AI in the Learning Process

A further challenge is identifying when assistance supports learning and when reliance on AI reduces students' participation in essential thinking and decision-making. In a survey of 300 university students, Zhang et al. (2024) found that the relationship between academic self-efficacy and AI dependency involved academic stress and performance expectations in the model they examined. This finding cautions against explaining AI dependency solely in terms of low self-efficacy. Because the data were cross-sectional, the direction of the relationship between self-efficacy and dependency remains unclear.

Whether students remain engaged in learning is better judged by whether they can check outputs, evaluate assistance, and justify their decisions than by how often they use AI. Findings on help-seeking and on essay improvement versus transfer support designing activities that preserve these student responsibilities (A. Chen et al., 2025; Fan et al., 2025).

Methodology

This technical article uses a purposive narrative synthesis to develop a learning design and assessment proposal. The starting set comprised the sources cited in the version first submitted; the search used to assemble them was not documented and cannot be reconstructed. The review therefore makes the role of each source explicit and does not claim a reproducible search. Sources serve three purposes: foundational self-efficacy theory and measurement; higher education studies measuring academic or task-specific self-efficacy in relation to Generative AI; and studies of AI-supported performance, transfer, or feedback evaluation. Table 1 summarizes every empirical study of the latter two groups cited in this article, recording their tasks, assistance, designs, outcomes, and inferential limits. Performance and process studies inform design without constituting evidence of self-efficacy change. The selection rationale and scientific reasoning are described with reference to narrative-review quality guidance (Baethge et al., 2019).

Carless and Boud (2018), Bandura (2006), and Baethge et al. (2019) were added during revision to support feedback uptake, measurement, and source-selection descriptions, respectively. Renkl and Atkinson (2003) and Schraw (2009) were added to support, respectively, the discussion of faded assistance and the distinction among calibration

measures; their bibliographic details and relevance to these purposes were checked against publisher or institutional records on 17 October 2026. The revision did not comprehensively update the empirical literature. No exhaustive database coverage, screening counts, or formal risk-of-bias assessment is claimed; the synthesis cannot estimate prevalence or pooled effects.

The analysis is organized around three dimensions. The first distinguishes academic self-efficacy, AI self-efficacy, and task performance to clarify what each study measures and which conclusions its findings support. The second examines evidence from studies of writing, programming, and feedback use in relation to the information learners use to judge their capabilities. Bandura's account of self-efficacy provides the theoretical basis for considering how learners interpret their contributions and the assistance they receive during task completion (Bandura, 1997, 2006; Usher et al., 2023). The third connects these interpretations to proposals for learning activities and assessment, including opportunities to evaluate AI feedback, explain decisions, and demonstrate performance under specified assistance conditions (Bearman et al., 2024; Fan et al., 2025).

Across these dimensions, the article distinguishes findings that directly concern self-efficacy from evidence about related learning processes and performance outcomes. The latter informs the design proposals without being treated as direct evidence of changes in academic self-efficacy. The proposed framework draws on both to specify activities, the evidence they produce, and how that evidence can be measured. Its relationships remain propositions for future investigation, and the educational effects of the framework require empirical evaluation.

Technical Proposals

Proposed Conceptual Framework

Figure 1 organizes four conceptual elements: (1) the target capability and task criteria, (2) permitted AI assistance, (3) evidence of learner decisions, and (4) outcomes assessed separately. Element 4 has two components: task-specific self-efficacy (4a) and demonstrated performance (4b); the labels do not denote two successive stages. Task context, prior knowledge, and task difficulty frame the specification of elements 1 and 2 and delimit how the later elements are interpreted. Performance experience and its interpretation describe the theoretical process through which learners may judge their own contribution. The diagram distinguishes theory-informed relationships from the author's proposed design links; none of its arrows represents a causal effect tested by this article.

The solid arrows between performance experience, its interpretation, and self-efficacy summarize Bandura's theoretical account of how interpreted experience informs capability beliefs (Bandura, 1997; Usher et al., 2023). They are not estimates of effects in AI-assisted learning, and the figure does not exhaust the four sources of self-efficacy. Dashed arrows identify the author's proposed connections between task specifications, assistance, decision evidence, and assessment. Evaluative judgement supports asking students to check AI suggestions and justify their use (Bearman et al., 2024), but whether these activities improve academic self-efficacy or its calibration remains to be tested. The arrows leading to the final comparison indicate an analytic procedure, not a causal influence of one outcome on the other.

The conceptual elements in Figure 1 identify what must be specified and observed; the five implementation phases described in the Implementation section identify when instructors and learners act. Specify establishes elements 1 and 2. Attempt, Evaluate, and Revise generate element 3 and opportunities to interpret performance experiences. Demonstrate collects element 4 through a separate no-AI assessment, while self-efficacy for the original assisted condition is measured at the start of Attempt and again at the end of Revise. Thus, the four-element framework and five-

phase sequence have different functions and need not have matching numbers. The independent task extends the assessment beyond the assisted activity without making its capability rating interchangeable with the earlier ratings (Bandura, 2006; Fan et al., 2025).

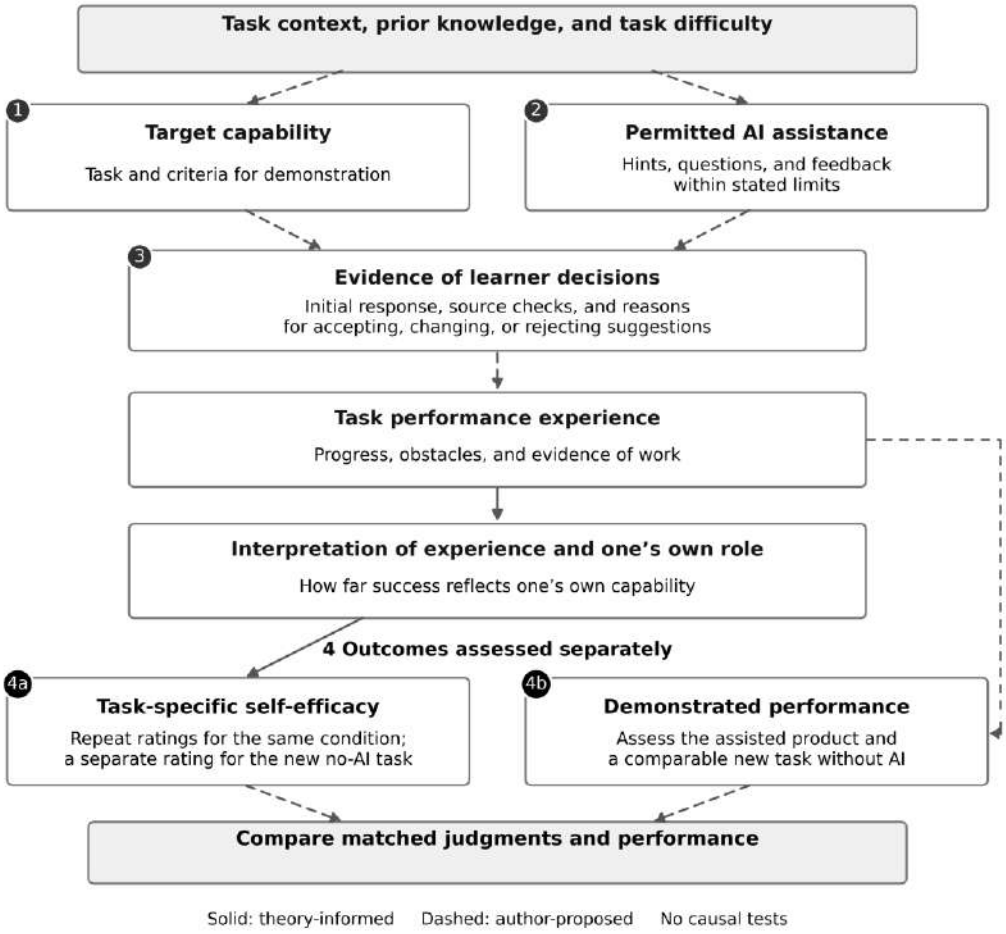


Figure 1 Conceptual elements and theory-informed and proposed links in AI-assisted learning

Application of the Sources of Self-Efficacy to Learning Design

The four sources of self-efficacy provide a way to examine the information students may use when judging their capabilities. Table 2 turns each source into a possible design feature and separates established theory from applications proposed for Generative AI contexts.

Table 2 Applying the sources of self-efficacy to Generative AI–assisted learning design

Theoretical source	Proposed learning design	Evidence and cautions
Mastery experiences	Offer hints or explanations while retaining a task that requires students to apply the target skill.	Record the assistance used and ask students to explain their solution. Use a task without AI when making a claim about independent performance.
Vicarious experiences	Compare AI-generated examples with examples of how people approach and revise a task.	Examine what students learn from each example; an AI-generated output should not be assumed to function like observing another learner.
Social persuasion	Use AI suggestions as material for discussion about progress and possible improvements.	Check the accuracy and specificity of suggestions and examine how students interpret their credibility. AI feedback should not be assumed to have the same effect as feedback from a person.
Physiological and affective states	Allow practice and clarification while preserving access to instructor support.	Examine anxiety and perceived capability separately from satisfaction with the tool or task.

Note. The table is the author's conceptual application of Bandura (1997) and Usher et al. (2023).

The proposed framework also identifies relationships for future empirical testing. The association between AI assistance and task-specific self-efficacy may vary with the decisions students make while completing a task. Explaining and checking feedback may affect how closely their capability judgments correspond to demonstrated performance, while prior knowledge and task difficulty may shape both relationships. Future studies could examine them by recording each student's assistance conditions and decisions and measuring self-efficacy and performance at multiple time points (A. Chen et al., 2025; Fan et al., 2025; Honicke et al., 2023; Yang et al., 2026).

Implementation

The activities and assessment procedures in this section are proposals derived from the framework. None has been empirically tested, and their effects on academic self-efficacy remain to be evaluated.

Learning Design Directions

Activity Design

Implementation begins by defining the target capability, task criteria, permitted assistance, and timing of AI access. Students rate their capability for the specified assisted condition before producing an initial response without AI. During practice, they may request bounded questions, hints, or feedback, check suggestions, and document consequential decisions. Immediately after revision, they repeat the identical capability items and instructions for the same task, criteria, sources, and assistance condition. The instructor examines the initial and revised products alongside the decision evidence. A later task without AI receives its own pre-task capability rating and separate performance assessment (Bandura, 2006; Bearman et al., 2024; A. Chen et al., 2025).

Order and Withdrawal of AI Assistance

The order of assistance is a design variable, not an invariant property of the framework. The worked example uses an independent attempt followed by bounded AI

feedback and a later independent task. This sequence preserves an initial response produced before AI exposure. An AI-first demonstration may instead be appropriate when learners lack prerequisite knowledge; research on worked examples and fading provides a rationale for testing that option, not proof of its effects with Generative AI (Renkl & Atkinson, 2003). If AI is used first, the subsequent response cannot be treated as an unexposed baseline; a diagnostic must precede AI access if such a baseline is required.

For an AI-first adaptation of Table 3, Specify establishes the task and assistance rules and collects any pre-exposure diagnostic. Attempt places a bounded AI demonstration before the learner-produced response. Evaluate checks the demonstration against sources and criteria; Revise requires a justified revision; Demonstrate retains the new no-AI task. Collect the first efficacy ratings at the start of Attempt, before the demonstration, and repeat them at the end of Revise under identical task and assistance specifications. The four conceptual elements remain unchanged. This adaptation requires empirical testing (Bandura, 2006; Renkl & Atkinson, 2003).

Alternating supported practice and independent attempts is another proposed implementation. Each episode should record when AI was available, what assistance was permitted, and which judgment and product it produced. Earlier assistance may influence later no-AI performance, which makes the latter evidence of performance after support, not a baseline or proof that a particular sequence caused learning. In every sequence, repeat ratings only within a fixed task and assistance condition and keep later independent-task judgments separate (Bandura, 2006; Fan et al., 2025).

Feedback Design

AI feedback should be directed at a defined feature of the work so that students can assess its relevance. In academic writing, a student might draft a claim and supporting reasons before asking AI to identify gaps in the argument. The student checks each suggestion against the criteria and supplied sources, then explains any revision. Carless and Boud (2018) describe feedback literacy as involving appreciation of feedback, judgment, management of affect, and action. The proposed record below focuses on judgment and action; it does not assess all aspects of feedback literacy. Its application to AI feedback also draws on evaluative judgement (Bearman et al., 2024).

A brief decision record contains five fields: (1) the AI suggestion, with the relevant interaction excerpt; (2) the decision to accept, adapt, or reject it; (3) the criterion and source passage used to check it; (4) the student's reason; and (5) the resulting change, linked to the initial and revised drafts. An unchanged draft can be accompanied by a justified rejection. This is a proposed documentation structure, not a validated instrument or evidence that reflection occurred (Bearman et al., 2024; Carless & Boud, 2018).

For a pilot, the instructor could review one consequential decision per student against the source passage and draft, requesting further records or explanation when discrepancies arise. This starting rule requires evaluation. Recording review time per student, follow-up cases, and unresolved discrepancies would test the workload proposal in the Discussion.

Assessment Design

Assessment distinguishes academic task self-efficacy, AI self-efficacy, and performance (Bandura, 2006; Y.-Y. Wang & Chuang, 2024). For the worked example, T0 is the capability rating before the initial response; T1 repeats the same items and instructions immediately after revision, with the task, source set, criteria, and permitted AI assistance held constant. T1 minus T0 describes change in the same specified capability judgment, although familiarity with the task and repeated measurement are alternative explanations to an

instructional effect. T2 is collected before the later, different task without AI and concerns independent capability for that task. T2 must not be subtracted from T0 or T1 and interpreted as change in the same judgment. To study change in independent self-efficacy, researchers need an additional baseline and follow-up instrument aligned to a prespecified family of comparable no-AI tasks.

Calibration analysis requires judgments and performance criteria that correspond. Schraw (2009) distinguishes absolute accuracy, relative accuracy, and bias; these distinctions inform monitoring questions but do not validate a self-efficacy instrument. The example below therefore includes separate capability items for claim clarity, source accuracy, and reasoning, aligned with Table 4. Report these ratings alongside the corresponding rubric categories, rather than assuming that a correlation establishes calibration. Numerical discrepancy scores require a defensible common metric. Without one, describe patterns such as high capability ratings accompanying weak performance; do not label a single discrepancy as established overconfidence. The later no-AI task provides a separate comparison under its stated condition (Bandura, 2006; Schraw, 2009).

Worked Example in Academic Writing

Consider a second-year educational technology course in which students write a 150–200-word argumentative paragraph on whether AI feedback tools should be used in secondary school writing instruction. The instructor supplies two short sources and criteria for claim clarity, source accuracy, and reasoning. Before the initial draft, students complete T0 for this task with AI questions permitted only after the draft. They then draft without AI. During revision, AI may question unclear claims or unsupported reasoning but may not supply a finished argument or new sources. Students check suggestions and record their decisions, then complete T1 using the identical T0 items and instructions. In a later session, they complete T2 for a new 150–200-word task on AI-generated quizzes in formative assessment and write without AI, using two new sources of comparable length and difficulty. Shared criteria support comparison, but task equivalence requires piloting and cannot be assumed from length alone. The writing products, decision evidence, and ratings at T0, T1, and T2 remain distinguishable observations (Bandura, 2006; Bearman et al., 2024; A. Chen et al., 2025).

For T0 and T1, use the shared instruction: ‘For Task A on AI feedback in secondary school writing, use the two supplied sources to write 150–200 words. AI may ask questions after your initial draft but may not write the argument or add sources. How certain are you that you can do each of the following?’ Rate each item from 0 (cannot do this at all) through 50 (moderately certain I can do this) to 100 (highly certain I can do this) (Bandura, 2006).

Claim clarity: Write a clear and consistent position that directly answers the prompt.

Source accuracy: Represent both supplied sources accurately and connect them to relevant statements.

Reasoning: Explain how the evidence supports the position and qualify the argument where necessary.

Repeat the identical instructions and all three items at T1. For T2, name Task B, its new topic and sources, and the condition ‘without AI at any stage’; retain the criterion wording. Examine each rating against its corresponding rubric criterion. These are author-designed item examples requiring content review and piloting; no validated total score or numerical calibration index is assumed.

Table 3 Proposed sequence and evidence for an AI-assisted academic writing task

Phase	Learner task and permitted AI assistance	Evidence and interpretation
Specify	Define Task A, two supplied sources, writing criteria, and AI questions allowed only after the initial draft.	Elements 1 and 2 establish a fixed referent for the repeated T0 and T1 ratings for each criterion.
Attempt	Complete T0 for Task A under the stated assisted condition, then draft independently before AI access.	T0 (4a) concerns the entire specified activity; the initial draft is a separate no-AI performance observation.
Evaluate	Use AI questions during revision and check suggestions against the sources and criteria.	Source checks and selected interaction excerpts document element 3 and feedback evaluation.
Revise	Revise and justify selected decisions; then repeat the identical T0 items and instructions as T1 for Task A under unchanged assistance rules.	The T0 and T1 ratings (4a) concern the same judgment. Draft comparison and decision notes provide distinct process and performance evidence.
Demonstrate	Complete T2 for new Task B under no-AI rules, then write independently with new sources and the same rubric.	T2 is a separate capability judgment (4a); Task B provides independent performance evidence (4b), not a causal test of transfer.

Table 4 provides provisional descriptors for both writing tasks. The rubric assesses products; explanations of AI decisions are reviewed separately as process evidence. Instructors should pilot the descriptors on examples, discuss disagreements between raters, and examine reliability before research use. Neither these descriptors nor the sample capability items constitute a validated instrument. A writing total of 0–6 must not be directly subtracted from a 0–100 self-efficacy rating (Bandura, 2006; Schraw, 2009).

Table 4 Provisional analytic rubric for Tasks A and B

Criterion	0 Not demonstrated	1 Partly demonstrated	2 Clearly demonstrated
Claim clarity	No identifiable position, or the position does not address the prompt.	A relevant position is identifiable but ambiguous or inconsistent.	A clear, consistent position directly answers the prompt.
Source accuracy	Sources are absent, fabricated, or materially misrepresented.	Source use is partly accurate but incomplete or includes a minor misrepresentation.	Both supplied sources are represented accurately and linked to relevant statements.
Reasoning	No explanation connects the evidence to the claim.	Some links are explained, but a key inference is unsupported.	The paragraph explains how relevant evidence supports the claim and qualifies it where needed.

Note. Rate each product criterion from 0 to 2 (total 0–6). Review decision explanations separately; these descriptors are proposed examples requiring validation.

Future Research Directions

Assistance Sequences and Levels

Future studies should compare assistance sequences as well as assistance levels. One design could randomly assign learners to an independent attempt followed by AI feedback, an AI example followed by faded assistance, or a clearly specified alternating sequence. Groups need comparable tasks, practice time, sources, and feedback opportunities. An independent diagnostic before exposure and a common final no-AI task would help separate initial differences from later performance. Practice, order, and carryover effects must be considered, especially when the same learners experience several sequences. The framework identifies these comparisons but does not predict a universally superior order (Fan et al., 2025; Renkl & Atkinson, 2003).

Correspondence Between Self-Efficacy and Performance

Studies of calibration should examine how students' judgments of their capabilities correspond to performance across tasks of specified difficulty and AI-use conditions. Measuring both over time would show whether growing self-efficacy accompanies demonstrated improvement, remains stable despite progress, or rises without comparable performance gains. Capability judgments should refer to the same task and assistance condition as the performance being assessed; otherwise, the correspondence between them is difficult to interpret. This approach reflects the task-specific nature of self-efficacy and allows researchers to examine how performance on one task may inform students' judgments about the next (Bandura, 2006; Honicke et al., 2023).

Human–AI Collaborative Learning

The proposed division of roles among students, instructors, and AI also requires testing. One design could compare AI hints during practice followed by instructor-guided reflection with assistance from a single source, while keeping task difficulty, learning time, and opportunities for feedback comparable. Outcomes should include students' evaluation of assistance, task-specific self-efficacy, and performance. Differences observed in help-seeking processes and in outcomes under teacher and AI assistance provide a rationale for testing whether a shared arrangement is superior (Bearman et al., 2024; A. Chen et al., 2025; Yang et al., 2026).

Long-Term Outcomes

Follow-up studies should examine whether changes in students' capability beliefs and performance persist beyond the activity in which Generative AI was used. Assessments at later points could ask students to complete a comparable new task, first under a clearly specified assistance condition and, where appropriate, with less or no AI support. Researchers should assess self-efficacy for each condition separately and distinguish performance on the original task from performance on the new one. This would help determine whether students can apply what they learned when the task or available support changes. Longer-term evidence is needed because it remains unknown whether the improvement in writing self-efficacy reported by Huang and Mizumoto (2024) was sustained over time.

Discussion

The framework's contribution is to connect instructional choices with interpretable evidence: task and assistance specifications establish what a capability judgment concerns, decision records document participation, and separate performance assessments show what students demonstrate. The contrasting outcomes in Yang et al. (2026) and Fan et al. (2025) motivate this separation. They do not establish that the proposed sequence will raise self-efficacy or improve calibration. Repeated ratings under a fixed condition can describe changes in perceived capability; explaining those changes requires an appropriate comparison design.

Implementation has practical limits. AI feedback may be inaccurate or misaligned with disciplinary criteria; instructors should provide authoritative source material, require source checks, and retain responsibility for resolving disputed suggestions (Bearman et al., 2024; Miao & Holmes, 2023). Reviewing complete interaction logs may also be impractical in large classes. A proposed compromise is a brief record of selected consequential decisions, checked against the draft and a relevant excerpt of the interaction, with targeted instructor review. This reduces the volume of records, but its workload and adequacy as evidence require evaluation; a fluent justification alone cannot establish that a learner carried out the reasoning.

Transfer across subjects also requires adaptation. The writing example is suited to tasks with explicit claims, accessible sources, and inspectable revisions. In programming, a justified debugging decision and a new no-AI coding problem could replace the paragraph; in quantitative subjects, a worked solution and explanation could serve analogous roles. These are proposed applications, not evidence of cross-disciplinary effectiveness. Instructors must redefine the target capability, permitted assistance, product criteria, and task difficulty for each discipline; the existing programming and writing findings do not establish equivalence across contexts (Huang & Mizumoto, 2024; Yilmaz & Karaoglan Yilmaz, 2023).

Conclusion

The framework specifies capabilities, AI assistance, learner decisions, and separate belief and performance outcomes through five instructional phases. It proposes repeated capability ratings under fixed conditions and a later no-AI assessment. Empirical studies should validate the measures, compare assistance sequences, assess instructor workload, and test disciplinary adaptations.

References

Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA—a scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review, 4*, Article 5. https://doi.org/10.1186/s41073-019-0064-8

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman. https://search.worldcat.org/title/36074515

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age Publishing.

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education, 49*(6), 893–905. https://doi.org/10.1080/02602938.2024.2335321

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

Chen, A., Xiang, M., Zhou, J., Jia, J., Shang, J., Li, X., Gašević, D., & Fan, Y. (2025). Unpacking help-seeking process through multimodal learning analytics: A comparative study of ChatGPT vs human expert. *Computers & Education, 226*, Article 105198. https://doi.org/10.1016/j.compedu.2024.105198

Chen, S., & Cheung, A. C. K. (2025). Effect of generative artificial intelligence on university students learning outcomes: A systematic review and meta-analysis. *Educational Research Review, 49*, Article 100737. https://doi.org/10.1016/j.edurev.2025.100737

Chen, X. (2026). Layer-sensitive cognitive offloading in generative AI-assisted writing: Supported performance and independent no-AI outcomes. *Frontiers in Psychology, 17*, Article 1906199. https://doi.org/10.3389/fpsyg.2026.1906199

Dai, Z. (2026). Rethinking AI-assisted writing instruction: Feedback literacy scripts, calibration training, and student writing development. *Frontiers in Psychology, 17*, Article 1829268. https://doi.org/10.3389/fpsyg.2026.1829268

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Honicke, T., Broadbent, J., & Fuller-Tyszkiewicz, M. (2023). The self-efficacy and academic performance reciprocal relationship: The influence of task difficulty and baseline achievement on learner trajectory. *Higher Education Research & Development, 42*(8), 1936–1953. https://doi.org/10.1080/07294360.2023.2197194

Huang, J., & Mizumoto, A. (2024). Examining the effect of generative AI on students' motivation and writing self-efficacy. *Digital Applied Linguistics, 1*, Article 102324. https://doi.org/10.29140/dal.v1.102324

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist, 38*(1), 15–22. https://doi.org/10.1207/S15326985EP3801_3

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3

Usher, E. L., Butz, A. R., Chen, X.-Y., Ford, C. J., Han, J., Mamaril, N. A., Morris, D. B., Peura, P., & Piercey, R. R. (2023). Supporting self-efficacy development from primary school to the professions: A guide for educators. *Theory Into Practice, 62*(3), 266–278. https://doi.org/10.1080/00405841.2023.2226559

Wang, Y. (2026). The effect of generative artificial intelligence literacy on academic achievement: The mediating role of academic self-efficacy and the moderating role of critical thinking. *Frontiers in Psychology, 17*, Article 1761562. https://doi.org/10.3389/fpsyg.2026.1761562

Wang, Y.-Y., & Chuang, Y.-W. (2024). Artificial intelligence self-efficacy: Scale development and validation. *Education and Information Technologies, 29*(4), 4785–4808. https://doi.org/10.1007/s10639-023-12015-w

Yang, X., Liu, X., & Gao, Y. (2026). The impact of generative AI on students' learning: A study of learning satisfaction, self-efficacy and learning outcomes. *Educational Technology Research and Development, 74*(1), 235–248. https://doi.org/10.1007/s11423-025-10540-8

Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence, 4*, Article 100147. https://doi.org/10.1016/j.caeai.2023.100147

Zhang, S., Zhao, X., Zhou, T., & Kim, J. H. (2024). Do you have AI dependency? The roles of academic self-efficacy, academic stress, and performance expectations on problematic AI usage behavior. *International Journal of Educational Technology in Higher Education, 21*, Article 34. https://doi.org/10.1186/s41239-024-00467-0

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology, 25*(1), 82–91. https://doi.org/10.1006/ceps.1999.1016