Scilight

*Library, Information & Services*

https://www.sciltp.com/journals/lis

*Perspective*

# Controlling the Instructional Structure in Generative AI Learning Evaluation Experiments

Zerong Xie * and Senze Gao

Shengda Future (Guangzhou) Consulting Co., Ltd., Guangzhou 510000, China
* Correspondence: hijetbrains@gmail.com

**How To Cite:** Xie, Z.; Gao, S. Controlling the Instructional Structure in Generative AI Learning Evaluation Experiments. *Library, Information & Services* **2026**, *1*(1), 6.

Received: 3 December 2025
Revised: 26 May 2026
Accepted: 28 May 2026
Published: 10 June 2026

**Abstract:** Comparisons of Generative AI (GenAI) learning interventions raise an important methodological consideration regarding instructional structure. The structural fluidity of AI interactions may introduce uncontrolled variability that confounds between-group comparisons. This is a sharp instance of a broader threat to validity in learning evaluation experiments, where the surrounding instructional structure often remains unspecified even when a particular research element is being tested. Confounders such as instructor delivery, the clarity of an opening explanation, or the form of practice may systematically bias inferences about the manipulated variable. This commentary proposes a four-phase framework (Establish Relevance, Technical Details, Intuition, and Practice) for aligning the instructional structure of learning evaluation experiments. Using the teaching of recursion in computer science as a case study, we demonstrate a procedure for standardizing each phase across conditions. We then examine how the probabilistic nature of large language models complicates, without invalidating, structural control in GenAI research, and we identify practical strategies for converting stochastic output variance from a structural confound into bounded measurement noise.

**Keywords:** educational technology; instructional design; artificial intelligence; generative AI; instructional fidelity; large language models; learning evaluation

## 1. Motivation

In discipline-based education research, the black-box nature of instructional interventions often complicates data interpretation. Valid comparisons of pedagogical approaches require evidence that observed effects are attributable to the method itself and not to confounding factors such as the instructor's enthusiasm or the clarity of the opening explanation [1]. The lack of a controlled instructional structure may produce low fidelity of implementation, in which the intervention delivered differs from the intervention as designed [2]. This concern is especially acute in GenAI educational research, where the structural fluidity of AI interactions introduces substantial variance unless it is bounded by a consistent pedagogical architecture [3]. We therefore propose a standardized instructional framework that serves as a substrate for experimental design. Students assigned to different conditions receive structurally identical instruction up until the point at which the manipulated variable is introduced. The remainder of this commentary delineates the four phases of the framework (Relevance, Technical Details, Intuition, and Practice) and then returns to GenAI to examine how the framework constrains, without eliminating, the variability inherent to large language models.

Recent syntheses likewise emphasize that GenAI studies require transparent reporting of model configuration, prompting, human oversight, data provenance, and instructional context so that observed effects are attributable to learning design and not opaque system behavior [3–5]. Meta-analytic and experimental evidence

further indicates that teacher scaffolding, learner agency, and task design moderate the benefits and risks of GenAI support [6–8].

## 2. A Four-Phase Instructional Framework

We illustrate the framework with a case study from computer science education, namely the teaching of recursion. Recursion is a useful test case because it is conceptually difficult, has well-documented pedagogical pitfalls [9], and admits both rigid and authentic forms of practice, all features that exercise each of the four phases below. The phases are intended to be controlled jointly. When each is held constant across conditions, except where the experimental manipulation explicitly varies it, researchers are positioned to isolate the contribution of the manipulated variable while preserving a consistent pedagogical arc.

Figure 1 visualizes this logic. The framework functions as a control template in which each phase is specified, documented, and held constant unless it is itself the planned manipulation.

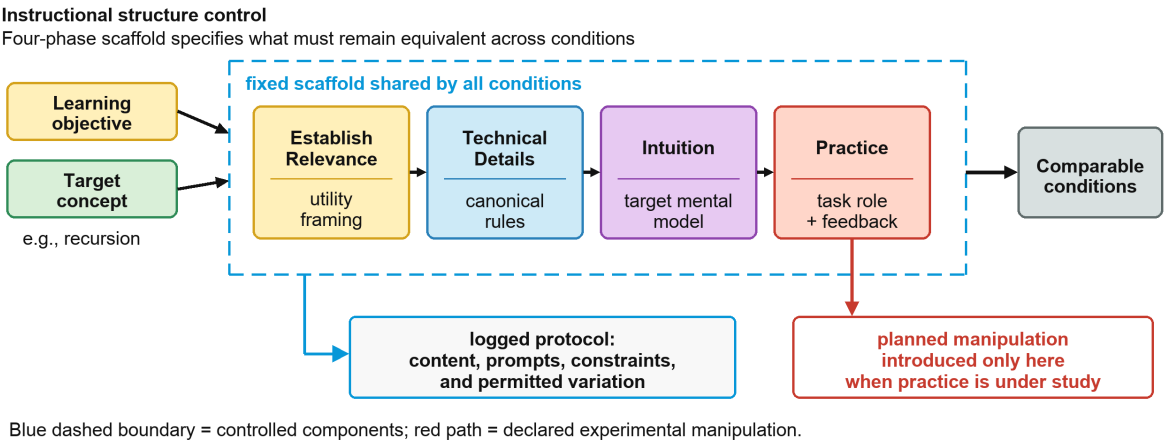


**Figure 1.** Four-phase instructional framework for controlling pedagogical structure across experimental conditions.

### *2.1. Establish Relevance*

The function of Establish Relevance is to demonstrate the necessity and utility of the concept before the instruction delves into the mechanics. Although early motivational design models such as Keller's ARCS framework [10] first foregrounded relevance as a precondition for engagement, contemporary educational psychology offers a richer account of how perceived relevance translates into sustained interest and effort. Expectancy-value research has shown that brief, well-designed interventions that help students articulate the personal utility of course content reliably increase situational interest, behavioral engagement, and downstream achievement [11,12]. Subsequent reviews have generalized this finding into a broader account of relevance for learning and motivation, in which relevance is not a single message delivered by an instructor but an ongoing meaning-making process negotiated between content, learner goals, and life context [13,14]. Closely related work on interest development further distinguishes triggered situational interest from the more durable, maintained interest that supports persistence in a domain [15]. Taken together, this literature reframes the Establish Relevance phase as a deliberate scaffolding move whose effectiveness depends on the alignment between the chosen utility cue and the learners' existing values; the mere presentation of a "real-world" example is insufficient.

In an experimental setting, this phase is therefore standardized with a more modest objective. Because students enter a classroom with substantial individual differences in prior knowledge, identity, and personal interest, no relevance script guarantees that all groups arrive at the experimental manipulation with equal motivation. The realistic aim is to establish a comparable baseline of perceived relevance across conditions, an equivalent floor of utility-value framing, thereby supporting more confident attribution of subsequent differences in engagement or learning to the experimental variable, not to asymmetric framing. Motivational equivalence is therefore treated as a controlled-for tendency, not a guaranteed state, and is ideally verified with a brief pre-task interest or utility-value measure.

In a typical computer science curriculum, students encounter recursion after mastering iteration. The instructional challenge is to justify why a new method is needed when loops already suffice for many tasks. A common but flawed approach is to introduce recursion through the factorial function. Mathematically, n! is expressible iteratively as $n \times (n - 1) \times \ldots \times 1$ or recursively as $n! = n \times (n - 1)!$ While these formulations are

mathematically equivalent, this example fails to establish relevance because the recursive implementation offers no clear advantage over the iterative one [16]. A stronger choice is to motivate recursion through problems that are difficult to solve iteratively but elegant when solved recursively, such as fractal geometry. Drawing a fractal tree using iterative loops requires complex stack management and advanced data-structure manipulation that exceeds the technical repertoire of most beginners. The recursive definition, by contrast, requires only that the program move forward, turn, and draw a smaller version of the tree [17]. This visual, structural example positions recursion as a tool for managing complexity instead of a mathematical curiosity, precisely the kind of utility-value linkage that recent relevance research identifies as motivationally productive. For example, a GenAI implementation of this phase prompts the tutor to render a fractal tree (either by generating Turtle Graphics code that students execute, or by producing a side-by-side comparison with the iterative alternative), accompanied by a fixed motivational script that frames the recursive solution as the elegant one. The relevance framing is specified in the system prompt or pre-task materials and is not allowed to emerge stochastically from each session, so that the artifact students see and the framing they read are identical across conditions, regardless of session-level variation in model response length or detail.

Recent commentary has argued that AI makes computational and mathematical ideas more meaningful when learners see how an abstract structure animates a concrete problem [18]. For the present framework, this reinforces the need to standardize the example, the representational form, and the motivational script used to make the example salient.

### *2.2. Technical Details*

The Technical Details phase corresponds to what implementation-fidelity research describes as the educative component of an intervention [2], namely the portion of the lesson in which the underlying rules of a system are made explicit. Rule comprehension is necessary for predicting the behavior of the artifact that students subsequently manipulate. Although our case study is drawn from computer science, the pedagogical move is not discipline-specific. At this point, the instructor externalizes the rules of a system precisely enough for independent learner reasoning. Shulman [19,20] framed this transformation of subject matter into teachable structure as the core of pedagogical content knowledge. The instructor's task is to identify the most powerful representations, analogies, and rule statements for a given concept, and to anticipate the preconceptions that will distort them. Without that explicit articulation, students often construct fragmented or surface-level mental models of the underlying mechanism [21].

The cross-disciplinary character of this phase becomes clearer with examples. In molecular biology, teaching the central dogma (DNA → RNA → protein) requires that the instructor commit to a canonical, syntactically precise account of transcription and translation, including the directionality of synthesis and the role of complementary base pairing, before students are asked to reason about mutations or gene regulation. Ambiguity at this stage produces durable misconceptions about, for example, the conditions under which a single nucleotide change propagates to a protein phenotype [22]. In literature, teaching synthesis across multiple sources similarly depends on a non-negotiable mechanical scaffold that distinguishes claims from evidence, treats citations as attribution, and uses transitions to signal logical relations beyond sequence. In the absence of that scaffold, student writing often remains at the level of serial summary instead of developing integrated arguments [23]. Across these cases, the pedagogy literature converges on a common prescription. Worked examples and explicit rule articulation are particularly effective when learners lack the schemas to chunk material themselves, because they reduce extraneous cognitive load and free working memory for schema construction [24]. Recursion is a convenient computer science instantiation of this general phenomenon.

For recursion specifically, this step requires defining the rigid structure of a recursive function. Regardless of whether students will later practice with mathematical or design-oriented tasks, all groups receive the same technical explanation. This involves articulating the execution flow (how a function creates a duplicate of itself and pauses its own execution until that duplicate completes) and defining the two non-negotiable components of the syntax, namely the Base Case and the Recursive Step. Standardizing this explanation prevents the confound in which one group understands the syntax better simply because it received a clearer lecture. Translating this requirement to a GenAI study means pinning the canonical exposition to a vetted source, typically through retrieval-augmented generation or a fixed reference document, so that the rule statement is not regenerated, and potentially altered, on each invocation. In practice, the experimenter instructs the model to deliver a scripted walkthrough of the Base Case and Recursive Step using a fixed example function, with the decoding temperature set near zero and a system prompt that anchors the wording, the worked example, and the order of presentation. If

students are permitted to ask follow-up questions, the model is constrained, via a guardrail prompt or a routing layer, to answer only from the canonical reference, so that the rule statements do not drift mid-session.

Consistent with recent reviews of LLMs in education, this documentation includes the model version, retrieval corpus, prompt constraints, and permitted follow-up scope, because the absence of these details limits replicability and makes it difficult to separate pedagogical effects from model behavior [3].

### *2.3. Intuition*

After absorbing the technical mechanics, learners often focus too narrowly on low-level details such as tracing execution line by line. The Intuition phase is intended to provide a higher-level thinking strategy or mental model that lets students apply knowledge without being overwhelmed by mechanical detail. The need for this phase is grounded in a robust finding from the cognitive science of expertise. Novices and experts represent the same problem differently. Novices tend to organize problems by surface features, whereas experts organize them around deep, principle-level structures that support transfer [25,26]. A purely mechanical exposition is associated with surface-level learner understanding. The Intuition phase explicitly seeds the deep-structure schema without relying on its emergence through extended practice.

This is not a discipline-specific challenge. Across fields, instruction supports movement of learners from the lower bands of Bloom's revised taxonomy (remembering and applying procedures) toward analyzing, evaluating, and creating, which require principle-level abstractions beyond rote execution [27]. The transition is difficult because the intrinsic cognitive load of holding a procedure in working memory crowds out the spare capacity needed to reflect on its structure. Cognitive load theory recommends managing this by offering schematic abstractions, worked examples, and analogies that compress the procedure into a single chunk [24]. The surface form of these strategies differs across disciplines (analogical mapping in physics, conceptual contrast cases in biology, argument schemas in writing), but each performs the same function of providing a portable mental model that survives changes in surface features.

In recursion, the critical intuitive leap is the Leap of Faith heuristic [28], which counters the well-documented tendency of novices to trace recursion all the way down to the base case, a process that overloads working memory [16]. The heuristic asks students to assume that the recursive call for a smaller input already works correctly, freeing them to focus solely on constructing the solution for the current input from that result. Standardizing this mental model across groups controls for differences in abstract reasoning strategies, ensuring that any observed performance differences reflect the experimental manipulation instead of differences in schema strength across groups. In a GenAI-mediated condition, the chosen mental model is specified explicitly in the system prompt. Otherwise, probabilistic generation may substitute a less powerful schema and introduce variance that is misclassified as individual differences. A concrete realization is to have the tutor present the Leap of Faith heuristic verbatim, accompanied by a fixed analogy (nested Russian dolls, for instance, or a manager who delegates a smaller version of the same task to a subordinate), and then to walk the student through a single worked example in which the heuristic is applied. Allowing the model to select analogies dynamically introduces analogy-level variance, including variation in explanatory fit and alignment with the intended heuristic.

### *2.4. Practice*

The Practice phase consolidates the learner's mental model through active application. Just as the previous phases establish the "why" and the "how," Practice is controlled to ensure that it aligns with the lesson's cognitive goals. In many experiments, this phase is treated as a generic activity (a uniform "do some problems" block), but a substantial body of cognitive psychology research shows that the form of practice has at least as much consequence for retention and transfer as its presence. Retrieval-based practice (the testing effect) reliably produces stronger long-term retention than equivalent time spent re-studying, particularly when the criterion test is delayed by days or weeks [29]. Practice that incorporates "desirable difficulties" such as spacing, interleaving across related problem types, and varied practice contexts slows apparent in-session improvement but substantially improves transfer to novel problems [30]. Conversely, massed, identical-format practice tends to inflate immediate performance while leaving learners poorly equipped to apply the concept in new settings. Treating Practice as undifferentiated therefore risks turning it into a hidden experimental variable. A practice block that is, in one condition, retrieval-heavy and interleaved and, in another, restudy-heavy and blocked is no longer a controlled comparison.

For these reasons, the framework requires that the nature of the practice be explicitly defined and standardized to prevent confounding variables such as task authenticity, retrieval demand, or student agency [31]. The framework distinguishes between rigid and authentic practice tasks in order to measure their specific effects. In a

rigid practice scenario, exemplified by the factorial problem, the solution path is convergent. A factorial calculation for 5 has a single correct result, 120, leaving no room for creative variation. An authentic practice scenario, such as using Turtle Graphics to generate fractals, involves divergent application in which the student acts as a designer [32]. Here, students generate new recursive patterns by altering branching angles or reduction ratios, fostering the deeper engagement that Turkle and Papert [33] termed epistemological pluralism. By characterizing the Practice phase at this level of detail, and by holding retrieval demand, spacing, and variability constant across conditions when they are not the variables under study, the comparison targets the intended pedagogical variables with reduced influence from accidental differences in task difficulty, retention demand, or engagement. GenAI implementations require analogous specification of retrieval demand, spacing, and feedback granularity in the prompt or pipeline, so that the form of practice does not drift toward the model's default response style. For a recursion study, this is operationalized in two contrasting ways. In the rigid condition, students submit factorial computations and the GenAI tool grades them against a fixed answer key, returning only correctness feedback. In the authentic condition, students are paired with the same model in a constrained "design partner" role, where it helps them invent novel fractal patterns by varying branching angles, recursion depth, or color schemes, but is instructed not to provide complete solutions. Specifying the model's role (grader versus collaborator), the feedback granularity, and the number of practice items protects the comparison from being silently reshaped by the model's default tendency to produce verbose, solution-revealing responses.

Recent empirical evidence supports treating Practice as a precisely specified intervention component. GenAI tools have been shown to support self-regulated science learning when embedded in a designed environment, yet they may also shift learners' agency, feedback behavior, or metacognitive effort depending on how the tool is framed and constrained [6,7,34,35]. The practice specification therefore includes the tool's role, feedback criteria, degree of solution disclosure, and whether learners first attempt retrieval or planning before receiving AI assistance.

## 3. Discussion and Conclusions

The case of generative AI illustrates a tension that has, until now, been largely implicit in instructional research. Conventional teacher-led instruction is variable in practice, but its variability is in principle bounded, in the sense that a lesson plan, a slide deck, or a script is available for inspection and reproduction. Large language models, by contrast, generate output by sampling from a probability distribution conditioned on a prompt and a (typically non-deterministic) decoding procedure. Their fluency is partly a property of this stochastic process, which Bender et al. [36] characterized as "stitching together sequences of linguistic forms… according to probabilistic information about how they combine," with no guarantee of semantic stability. In a classroom demonstration, two students issuing the same prompt may receive materially different explanations, examples, or even factual claims. In a between-groups experiment, nominal "exposure to a GenAI tutor" is therefore not a single, well-defined treatment [37].

Whether this uncontrolled variability matters for cross-group comparisons depends on the level at which the comparison is framed. If the research question is "does GenAI tool X improve learning relative to condition Y," then probabilistic output variance is part of the treatment, much as instructor-to-instructor variance is part of any human-led condition. The appropriate response is adequate sampling and replication, not the elimination of variance. If, however, the question concerns the effect of a specific pedagogical move (a particular explanation, analogy, or scaffold), then the probabilistic substrate becomes a serious confound, because the construct under study is not stably realized across participants. The framework proposed here offers a partial remedy. By specifying the required content of each of the four phases, researchers constrain the GenAI tool to operate within a defined pedagogical scope and avoid evaluating it as an unbounded conversational partner. Practical implementations include scripted system prompts that fix the relevance framing, retrieval-augmented constraints that pin technical details to a vetted source, explicit specification of the target mental model, and rubric-based filters on practice tasks. These measures do not eliminate stochasticity, but they reclassify it from a structural confound into bounded measurement noise, the same status it occupies in well-controlled human-instructor studies. Figure 2 summarizes this distinction by separating tool-level evaluations from mechanism-level evaluations.

Current evidence supports this distinction. GenAI interventions show positive average effects, with larger benefits observed when teacher support and task design are explicitly incorporated into the intervention [4,8]. Moreover, discipline-specific analyses suggest that students interpret AI assistance through different epistemic and professional concerns, which makes an explicit comparison structure necessary when studies cross domains or instructional contexts [38].

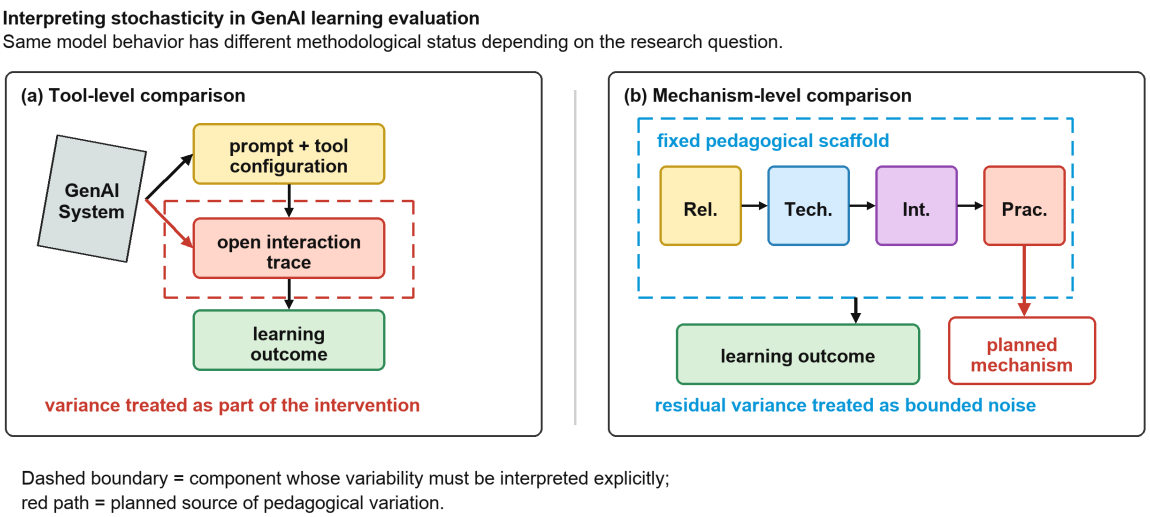


**Figure 2.** Evaluation logic for GenAI studies: stochastic output is interpreted differently depending on whether the study estimates tool-level effects or mechanism-level pedagogical effects. Note: Rel. = Establish Relevance; Tech. = Technical Details; Int. = Intuition; Prac. = Practice.

Standardizing the instructional structure is a precondition for credible learning evaluation, and the four-phase framework operationalizes that goal in a way that is portable across disciplines and across instructional media. Its added value in the GenAI era lies in making the pedagogical scaffold within which probabilistic model output operates explicit and inspectable. Ensuring that every phase (including Practice, the phase most often left undifferentiated) is deliberately designed and controlled, and that residual stochasticity is explicitly accounted for, supports movement in the field toward more reproducible, scientifically valid findings. Accordingly, the reporting specification for future GenAI learning-evaluation studies includes the pedagogical role of the model within each phase, the source of any fixed instructional content, the degree of permitted interactional variability, and the rationale for treating residual stochasticity as either treatment variance or measurement noise.

## Author Contributions

Z.X.: conceptualization, writing—original draft preparation; S.G.: conceptualization. All authors have read and agreed to the published version of the manuscript.

## Funding

This research received no external funding.

## Institutional Review Board Statement

Not applicable.

## Informed Consent Statement

Not applicable.

## Data Availability Statement

No new data were generated or analyzed in this study.

## Conflicts of Interest

The authors declare no conflict of interest.

## Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Gemini to check and fix grammar errors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## References


1. Taber, K.S. Experimental research into teaching innovations: Responding to methodological and ethical challenges. *Stud. Sci. Educ.* **2019**, *55*, 69–119. https://doi.org/10.1080/03057267.2019.1658058.
2. Century, J.; Rudnick, M.; Freeman, C. A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *Am. J. Eval.* **2010**, *31*, 199–218. https://doi.org/10.1177/1098214010366173.
3. Yan, L.; Sha, L.; Zhao, L.; et al. Practical and ethical challenges of large language models in education: A systematic scoping review. *Br. J. Educ. Technol.* **2024**, *55*, 90–112. https://doi.org/10.1111/bjet.13370.
4. Yan, L.; Greiff, S.; Teuber, Z.; et al. Promises and challenges of generative artificial intelligence for human learning. *Nat. Hum. Behav.* **2024**, *8*, 1839–1850. https://doi.org/10.1038/s41562-024-02004-5.
5. Yusuf, H.; Money, A.; Daylamani-Zad, D. Pedagogical AI conversational agents in higher education: A conceptual framework and survey of the state of the art. *Educ. Technol. Res. Dev.* **2025**, *73*, 815–874. https://doi.org/10.1007/s11423-025-10447-4.
6. Darvishi, A.; Khosravi, H.; Sadiq, S.; et al. Impact of AI assistance on student agency. *Comput. Educ.* **2024**, *210*, 104967. https://doi.org/10.1016/j.compedu.2023.104967.
7. Fan, Y.; Tang, L.; Le, H.; et al. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.* **2025**, *56*, 489–530. https://doi.org/10.1111/bjet.13544.
8. Gu, J.; Yan, Z. Effects of GenAI interventions on student academic performance: A meta-analysis. *J. Educ. Comput. Res.* **2025**, *63*, 1460–1492. https://doi.org/10.1177/07356331251349620.
9. Zmuda, M.A.; Hatch, M. Scheduling topics for improved student comprehension of recursion. *Comput. Educ.* **2007**, *48*, 318–328. https://doi.org/10.1016/j.compedu.2005.02.003.
10. Keller, J.M. Development and use of the ARCS model of instructional design. *J. Instr. Dev.* **1987**, *10*, 2–10. https://doi.org/10.1007/BF02905780.
11. Hulleman, C.S.; Harackiewicz, J.M. Promoting interest and performance in high school science classes. *Science* **2009**, *326*, 1410–1412. https://doi.org/10.1126/science.1177067.
12. Hulleman, C.S.; Godes, O.; Hendricks, B.L.; et al. Enhancing interest and performance with a utility value intervention. *J. Educ. Psychol.* **2010**, *102*, 880–895. https://doi.org/10.1037/a0019506.
13. Priniski, S.J.; Hecht, C.A.; Harackiewicz, J.M. Making learning personally meaningful: A new framework for relevance research. *J. Exp. Educ.* **2018**, *86*, 11–29. https://doi.org/10.1080/00220973.2017.1380589.
14. Harackiewicz, J.M.; Priniski, S.J. Improving student outcomes in higher education: The science of targeted intervention. *Annu. Rev. Psychol.* **2018**, *69*, 409–435. https://doi.org/10.1146/annurev-psych-122216-011725.
15. Renninger, K.A.; Hidi, S.E. *The Power of Interest for Motivation and Engagement*; Routledge: Oxfordshire, UK, 2016. https://doi.org/10.4324/9781315771045.
16. Bandi, A.; Fellah, A. The essence of recursion: Reduction, delegation, and visualization. *J. Comput. Sci. Coll.* **2018**, *33*, 115–123.
17. Gordon, A. Teaching recursion using recursively-generated geometric designs. *J. Comput. Sci. Coll.* **2006**, *22*, 124–130.
18. Xie, Z.; Zhang, C. Use AI in the classroom to bring problems to life. *Nature* **2025**, *644*, 338. https://doi.org/10.1038/d41586-025-02571-1.
19. Shulman, L.S. Those who understand: Knowledge growth in teaching. *Educ. Res.* **1986**, *15*, 4–14. https://doi.org/10.3102/0013189X015002004.
20. Shulman, L.S. Knowledge and teaching: Foundations of the new reform. *Harv. Educ. Rev.* **1987**, *57*, 1–23. https://doi.org/10.17763/haer.57.1.j463w79r56455411.
21. National Research Council. *How People Learn: Brain, Mind, Experience, and School*, Expanded ed.; National Academies Press: Washington, DC, USA, 2000. https://doi.org/10.17226/9853.
22. Wright, L.K.; Fisk, J.N.; Newman, D.L. DNA → RNA: What do students think the arrow means? *CBE-Life Sci. Educ.* **2014**, *13*, 338–348. https://doi.org/10.1187/cbe.CBE-13-09-0188.
23. Doolan, S.M. An exploratory analysis of source integration in post-secondary L1 and L2 source-based writing. *Engl. Specif. Purp.* **2021**, *62*, 128–141. https://doi.org/10.1016/j.esp.2021.01.003.
24. Sweller, J.; van Merriënboer, J.J.G.; Paas, F. Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* **2019**, *31*, 261–292. https://doi.org/10.1007/s10648-019-09465-5.
25. Chi, M.T.H.; Feltovich, P.J.; Glaser, R. Categorization and representation of physics problems by experts and novices. *Cogn. Sci.* **1981**, *5*, 121–152. https://doi.org/10.1207/s15516709cog0502_2.
26. Chi, M.T.H.; Glaser, R.; Farr, M.J. (Eds.). *The Nature of Expertise*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1988. https://doi.org/10.4324/9781315799681.
27. Krathwohl, D.R. A revision of Bloom's taxonomy: An overview. *Theory Into Pract.* **2002**, *41*, 212–218. https://doi.org/10.1207/s15430421tip4104_2.

28. Roberts, E. *Thinking Recursively*; John Wiley & Sons: Hoboken, NJ, USA, 1986.
29. Roediger, H.L.; Karpicke, J.D. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychol. Sci.* **2006**, *17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x.
30. Bjork, E.L.; Bjork, R.A. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*; Gernsbacher, M.A., Pew, R.W., Hough, L.M., et al., Eds.; Worth Publishers: New York, NY, USA, 2011; pp. 56–64.
31. Gulikers, J.T.; Bastiaens, T.J.; Kirschner, P.A. A five-dimensional framework for authentic assessment. *Educ. Technol. Res. Dev.* **2004**, *52*, 67–86. https://doi.org/10.1007/BF02504676.
32. Chiodini, L.; Sorva, J.; Hellas, A.; et al. Two approaches for programming education in the domain of graphics: An experiment. *Art Sci. Eng. Program.* **2025**, *10*, 14. https://doi.org/10.22152/programming-journal.org/2025/10/14.
33. Turkle, S.; Papert, S. Epistemological pluralism: Styles and voices within the computer culture. *Signs J. Women Cult. Soc.* **1990**, *16*, 128–157. https://doi.org/10.1086/494648.
34. Ng, D.T.K.; Tan, C.W.; Leung, J.K.L. Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *Br. J. Educ. Technol.* **2024**, *55*, 1328–1353. https://doi.org/10.1111/bjet.13454.
35. Steiss, J.; Tate, T.P.; Graham, S.; et al. Comparing the quality of human and ChatGPT feedback of students' writing. *Learn. Instr.* **2024**, *91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894.
36. Bender, E.M.; Gebru, T.; McMillan-Major, A.; et al. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual, 3–10 March 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 610–623. https://doi.org/10.1145/3442188.3445922.
37. Kasneci, E.; Sessler, K.; Küchemann, S.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274.
38. Yan, L.; Wang, H.; Xie, Z.; et al. The impact of artificial intelligence systems and tools on education: Comparative social media analytics of computing versus business students. *Systems* **2026**, *14*, 451. https://doi.org/10.3390/systems14040451.