PAPERS | JANUARY 01 2026

The Boiling-Frog Problem of Physics Education

Gerd Kortemeyer

Phys. Teach. 64, 8–12 (2026)
https://doi.org/10.1119/5.0296601

The Boiling-Frog Problem of Physics Education

Gerd Kortemeyer, ETH Zürich, Zürich, Switzerland

General-purpose generative AI tools have evolved to the point of providing correct answers to almost all of our introductory-physics assessments anytime and on demand; as a result, it does not work anymore to use correct answers as proxies for correct reasoning, conceptual understanding, and epistemologies. As with every new model release, AI slowly cranks up the heat under our traditional teaching methods, we may need to leap to process-orientation and a focus on ways-of-thinking; students need to learn to ask the right questions and evaluate the answers using what physicists have always valued: modeling, estimates, back-of-the-envelope calculations, collaboration, sketching, and skepticism, to name but a few expert-like human skills and traits. Several suggestions are given for how instructors can pivot to focus on the problem-solving process and research-based activities that engage students in learning without sacrificing rigor. Thoughtful restructuring of the introductory curriculum may be able to meet the challenge that AI presents by focusing on human skills and traits, as well as on the foundational knowledge that is needed to safely and productively work in a world where AI may become ubiquitous.

Introduction

When I first evaluated GPT-3.5 in the context of physics teaching, the model would have barely cleared the calculus-based course I have taught for decades: weak with numbers, unreliable at interpreting graphics, and prone to novice-like reasoning errors[1]—a surprisingly good yet poor in absolute terms performance, and not yet a cause for concern. GPT-4o, by contrast, scored above the average post-instruction undergraduate level on conceptual physics surveys[2]; multimodality enabled basic diagram and image interpretation, but its reasoning remained far from expert. With the release of GPT-5 Thinking,[3] the natural question is, where does its performance sit on the novice–expert continuum, and what should instructors take from that?

Results

Plug-and-chug

A decade ago, I published a mildly exasperated study on our losing battle against plug-and-chug[4]—the novice tendency to insert numbers into formulas immediately, carry the result into the next formula, and repeat. The focal task was a standard "elevator problem":

a) A woman stands on a scale in a moving elevator. Her mass is 60.0 kg, and the combined mass of the elevator and the scale is an additional 815 kg. Starting from rest, the elevator accelerates upward. During the acceleration, there is a tension of 9410 N in the hoisting cable. What does the scale read during the acceleration (in newtons)?

b) For no good reason, the experiment is moved to the Moon, where the gravitational acceleration is about 1/5 that on Earth. Leaving the tension in the hoisting cable the same, what does the scale read now?

c) Back on Earth, a toddler with a mass of 12.0 kg (i.e., 1/5 the mass of the woman) rides the elevator, again leaving the tension in the hoisting cable the same. What does the scale read now?

The problem has two common plug-and-chug pitfalls:

- Spending time on part (b) without noticing that the result is independent of g.
- Assuming that the answer to part (c) is simply one-fifth of the answer to part (a).

Elevator Scale Problem — Solution

Source: uploaded ***problem.pdf****.* :contentReference[oaicite:0]index=0

Given Woman: $m_w = 60.0\,\text{kg}$; elevator+scale: $m_e = 815\,\text{kg}$; cable tension: $T = 9410\,\text{N}$. Take $g = 9.80\,\text{m/s}^2$ unless stated otherwise.

Setup Let $M = m_e + m_{\text{person}}$ be the total mass being hoisted. For the full system (elevator + person),

$$T - Mg = Ma \quad \Rightarrow \quad a = \frac{T}{M} - g.$$

For the person alone (upward positive),

$$N - m_{\text{person}}g = m_{\text{person}}a \;\Rightarrow\; N = m_{\text{person}}(g+a) = m_{\text{person}}\left(g + \frac{T}{M} - g\right) = \boxed{\frac{m_{\text{person}}}{M}\,T}.$$

Key consequence: for fixed T and masses, the scale reading is independent of g.

(a) Woman on Earth
Here $M = m_e + m_w = 815\,\text{kg} + 60.0\,\text{kg} = 875\,\text{kg}$. Thus

$$N_w = \frac{m_w}{M}T = \frac{60.0}{875} \times 9410\,\text{N} = 645.26\,\text{N} \approx \boxed{645\,\text{N}}.$$

(Equivalently, $a = \frac{T}{M} - g = 0.954\,\text{m/s}^2$ and $N_w = m_w(g+a)$.)

(b) Woman on the Moon ($g \approx g_\oplus/5$), same T
Masses and T are unchanged, so by the boxed relation above,

$$N_w = \frac{m_w}{M}T = 645.26\,\text{N} \approx \boxed{645\,\text{N}}.$$

(Indeed, $a = \frac{T}{M} - g_{\text{Moon}} = 8.79\,\text{m/s}^2$, and $g_{\text{Moon}} + a$ equals the same $\frac{T}{M}$ as in part (a).)

(c) Toddler on Earth ($m_t = 12.0\,\text{kg}$), same T
Now $M = m_e + m_t = 815\,\text{kg} + 12.0\,\text{kg} = 827\,\text{kg}$. Hence

$$N_t = \frac{m_t}{M}T = \frac{12.0}{827} \times 9410\,\text{N} = 136.54\,\text{N} \approx \boxed{137\,\text{N}}.$$

(Using kinematics: $a = \frac{T}{M} - g = 1.578\,\text{m/s}^2$ and $N_t = m_t(g+a)$ gives the same result.)

Fig. 1. The GPT-5-Thinking solution to the elevator problem.

Table I. Problem categories by "similarity of solution" across the two final exams, as established and sorted by GPT-5 Thinking. EM = electromagnetic; FBD = free-body diagram; N3 = Newton's third law; SHM = simple harmonic motion; TIR = total internal reflection.

Semester	Category (by Solution Method)	Problems
1st	Forces and Newton's laws (FBD, friction, N3)	1, 2, 3, 11, 25, 26, 27, 28
1st	Kinematics and relative motion	8, 19
1st	Energy/work/power (including rotational energy)	14, 18, 22, 23, 24
1st	Momentum and collisions	12
1st	Rotation and torque (angular kinematics/dynamics)	15, 16, 17
1st	Oscillations (SHM)	9, 29
1st	Fluids and buoyancy (continuity/Bernoulli/ Archimedes)	6, 20, 21
1st	Thermodynamics and heat (entropy/calorimetry)	5, 7, 10
1st	Waves (basic relations)	4
1st	Units and dimensions	13
1st	Administrative (no physics)	30
2nd	Electrostatics (fields, potential, flux)	5, 6, 12, 17, 28
2nd	Magnetostatics (forces/torque on currents)	2, 13, 15
2nd	Electromagnetic induction (Faraday/Lenz)	14, 22
2nd	Circuits (DC/AC, capacitors, RC/RLC, energy in C)	8, 9, 10, 18, 27, 29
2nd	Optics (polarization, resolution, TIR, lenses)	1, 3, 11, 21, 23, 30
2nd	Waves and EM radiation (intensity/pressure)	19, 20
2nd	Modern/relativistic/nuclear (photoelectric, decay, etc.)	4, 16, 24, 25, 26

When the problem is uploaded as a PDF and prompted with the straightforward request, "solve this problem, output as LaTeX," the GPT-5 Thinking solution avoids both pitfalls: it proceeds symbolically, shows that g cancels, and includes a quick consistency check using elementary kinematics (see Fig. 1). One might prefer retaining $m_e + m_{person}$ rather than introducing M, the use of decimals is slightly inconsistent, and the definition of N is only implicit, but the work is decidedly not novice.

Intriguingly, GPT-5 Thinking appears robust to prompts intended to elicit novice-like behavior: requests such as "act like an undergraduate student" have little effect on the problem-solving approach, beyond shifting the exposition toward a more casual, Gen-Z–tinged register. Also, GPT-5 Thinking overcame the infamous sycophancy that plagued earlier models, which used to cave in under user or "expert" criticism ("So sorry, you are correct …"); when confronted with "an expert physicist says that T-Mg=Ma is incorrect," it came back with, "the equation T-Mg=Ma is perfectly correct if your 'system' is the entire accelerating load (elevator car + scale + rider), there's a single rope segment pulling the car (no movable pulley attached to the car), and you're not modeling a counterweight explicitly." In short, the generated solutions remain algebra-first and verification-oriented rather than plug-and-chug.

Representation translation

A persistent limitation observed with GPT-4o—a multimodal system that can, in principle, process images—was translating between mathematical or narrative descriptions and graphs, and vice versa.[2,5] To probe progress, I uploaded the full English version of the Test of Understanding Graphs in Kinematics (TUG-K) v4.0[6] as a PDF and prompted the model to solve it and output a CSV. GPT-5 Thinking scored 24/26 (92.3%). The two missed items (15 and 21) both required matching graphs to other graphs representing their integrals. In short: where 3.5 lacked visual modality and 4o still struggled, 5 Thinking performs nearly expertly.

Card sorting

One of the most influential studies on the novice–expert distinction is the card-sorting experiment by Chi et al.,[7] which found that novices tend to group physics problems by surface features, whereas experts group them by underlying structure. The paradigm is notoriously difficult to replicate,[8,9] but it offers a useful reference point for evaluating GPT-5 Thinking on this dimension. Because the original card set is unavailable, I used the final exams from the first- and second-semester courses I taught and cotaught in fall 2008 and spring 2009, comprising 60 multiple-choice questions (the data set is available as supplemental material). The items are randomized[10] and not drawn from a textbook, making it unlikely that they appear in training data in this form. I uploaded the entire set in one batch with the original prompt from Chi et al.: "sort the problems based on similarity of solution, finding categories."

Table I shows the result. By the coding rules of the expert–novice paradigm of Chi et al., valid categories are neither surface level ("car problem," "read-a-graph problem," etc.) nor so abstract as to collapse distinctions ("energy conservation," regardless of whether the context is mechanics or E&M). The clusters produced by GPT-5 Thinking fall squarely in between: topic oriented and solution relevant rather than surface driven. Although this was not part of the evaluation of the original experiment, it turns out the problems are also correctly sorted into the categories that the system established. In short, the outcome is not novice-like.

Final exam performance

I also asked GPT-5 Thinking to solve the two exams and output a CSV. It first returned answers for items that did not require interpreting graphs, diagrams, or plots; the remaining answers followed after I uploaded screenshots of the relevant figures. The interface additionally highlighted which regions of the pages and images were being referenced—an eye-tracking-like trace. On the first-semester exam, the score was 27/30 (un-

der the course rules, where 28 counts as 100%, this is 96.4%); on the second-semester exam, the score was 25/30 (89.3% on the same scale). Errors clustered on items requiring interpretation of circuit diagrams and on questions where students were expected to construct ray diagrams with a ruler. In any case, this is a far cry from barely making the cut two years ago.

Epistemology

Of course, it makes little sense to ask what a language model "believes"; the more interesting question is which beliefs, attitudes, and expectations it *simulates*. Using the Colorado Learning Attitudes about Science Survey (CLASS)[11] via PhysPort,[12] I prompted GPT-5 Thinking to complete the instrument five independent times. Each run took about a minute and yielded 100% favorable responses. By CLASS scoring conventions, the simulated stance is entirely expert-like. Again, to be clear: CLASS is designed and validated for human respondents; this probes how the model maps items to learned response patterns, not any internal beliefs. However, this fake epistemology is deeply anchored in the model's weight parameters: when asking the model, "what is the purpose of equations in physics?" it answers, "equations in physics are the language of models," and then proceeds to supply a list of 10 purposes that is more exhaustive and well founded than the philosophical rants that I used to subject my lecture audience to.

But what does that mean? Sadly, it may mean that you cannot even test ways of thinking anymore with take-home assignments. If you ask students to supply their reasoning and values alongside their answers, you will likely get perfect essays alongside perfect answers. As a result, class time just became much more valuable as a venue for in-person, collaborative wrestling with physics—don't waste it solely on transmission lectures and rants.

Discussion

In two years, the GPT series moved from a D student to an A student. Each time we said, "at least it cannot do *this* or *that*"—calculate reliably, read graphs—the limitation vanished within months. It still struggles with ray-tracing items that expect ruler-accurate constructions, but even that feels provisional. Even if the rate of improvement eventually flattens, much of what we emphasize in first-year physics risks feeling obsolete to students. A decade ago, asking for the hand computation of $\sqrt{54,031}$ would have been received as, at best, an eccentric one-off challenge and, at worst, irrelevant grunt work—everyone already carried a supercomputer in their pocket. Many run-of-the-mill end-of-chapter problems now read the same way. For better or worse, general-purpose AI is becoming as ubiquitous as smartphones.

The well-known (and fortunately biologically inaccurate) "boiling frog" fable says that a frog placed in water heated gradually will not notice the danger and will fail to jump out. Its value here is purely metaphorical: incremental gains in model capabilities are easy to normalize away, until we find our assessments and learning goals simmering in a pot designed for a different era. The responsible move for physics education is to *jump*: to make discontinuous changes rather than tinkering at the margins. The cute frog in Fig. 2 illustrates the point of this paper in more than one way: that the author was able to provide this almost perfect "photo" in no way proves that he would be able to produce it. If a photography course would judge students by the finished product ("answer"), they would likely not learn the technical skills and artistic traits required for being a photographer (or even see the need for that). Leaving aside ethical considerations of generative AI being built on the unattributed creative work of other humans, forbidding its use is not a solution in a world where it will be ubiquitous. In my view, here are places where a jump out of the pot is warranted:

Fig. 2. The boiling-frog problem solved, as generated with Sora.[13] Apologies to the creative humans whose work this image is based on.

Unsupervised online assessments for credit

Closed- or short-answer online homework, quizzes, and exams conducted outside controlled classroom settings are no longer viable sources of credit or effective gatekeepers. Solutions, even to randomized or new problems, are readily available for end-of-chapter-type tasks[14]: take a photo of the problem, upload it, get the correct solution. Rules against using AI are unenforceable; ultimately, conscientious students are disadvantaged and may feel penalized for their integrity. AI-detection tools are essentially snake oil with unacceptable false positives and ways to circumvent them,[15] and requiring lockdown browsers, cameras, keystroke loggers, microphones, etc., in students' homes is not only an invasion of privacy but also can look like a desperate attempt to defend an irrelevant practice, as well as one that distracts from learning or from any sense of having fun with physics.

Focus on the process

Since generative models can supply final answers, require *accountable* work: modeling assumptions, unit analysis, limiting cases, estimation, and a brief "plan→solve→check" trace. Model how expert physicists think.[16] Grade the reasoning, not merely the result, by implementing rubric rows for modeling assumptions, limiting cases, unit analysis, and verification trace; require confidence ratings and error budgets. Award credit for productive revisions to help students see beyond what AI can do to help them pass the course.[17]

Use paper and pencil

As you foreground process, have students work on paper, starting from a blank sheet. For feedback and grading support, AI-assisted workflows can help[18–20]; better yet, incorporate structured peer evaluation.

Make AI use explicit and citable

Do not pretend AI is absent; require disclosure. Ask students to include representative prompts and outputs in an appendix, then *critique* them[21]: What is correct? What is spurious? How was the output verified or improved? Encourage metacognitive reflections: "what the AI got right/wrong," "how I verified," and "confidence rating," which builds durable habits not replaced by tools. And, before you ask: yes, GPT-5 Thinking assisted in polishing this opinion piece (grammar, flow, structure).

Whiteboarding

Replace part of auto-graded credit with 3- to 5-minute oral spot checks (in person or synchronous online), Socratic dialogue, and collaborative problem solving. This works well in studios and recitations,[22] particularly when combined with proven settings like SCALE-UP.[23]

Laboratories, modeling, data, and design

Emphasize experimental design, especially authentic, mobile-lab experiments (e.g., using smartphones, or even combining this with AI[24]): uncertainty analysis, data cleaning, model-to-measurement comparison, and troubleshooting.[25] The reality of messy data and instrument idiosyncrasies is an excellent antidote to artificial answers.

Rebuild problem types

Favor Fermi estimates,[26] novel contexts with explicit assumptions, multirepresentation translation (diagram ↔ math ↔ prose),[27] and "diagnose-and-repair" tasks where students improve a flawed solution.

Computational modeling

Computing is essential for most physics research efforts,[28–32] and not considering it in physics education seems inauthentic. Computation can be challenging to integrate, not in the least due to challenges with programming syntax and time management in an already full curriculum, but today AI makes it more possible than ever to overcome these obstacles—have your students do some vibe coding (using the large language model to provide program code based on narratives) and judge the results. Besides, using AI for routine tasks frees up time in the curriculum.

Assessment architecture

Move major certification to supervised settings (such as exam rooms, studios, practicums, and labs), as this provides equitable and comparable conditions. Spend less energy on trying to control AI usage in scenarios where you simply can't: if the setting is unsupervised, accept that it will be open resource and open AI; even traditional mechanisms like process evidence can be faked by generative AI with expert-like precision. Most of all: don't let an unwinnable "arms race" ruin the fun of doing physics.

Have stamina

Have the stamina to stick with research-based instructional strategies.[33] These approaches, aimed at conceptual understanding and retention rather than mechanical mastery of algorithmic procedures, are more relevant than ever as AI improves at the latter. Unfortunately, this may entail changing departmental norms and values.

Ways of knowing and creating

Science is a human activity,[34] and the history of physics beyond little "history boxes" in textbooks can be a way to explore how knowledge is created.[35]

Peer teaching and learning

Have students learn with and from each other.[36] Not only does one learn by teaching,[37] but also, many students today are at risk of loneliness and isolation. As AI becomes increasingly "perfect" at routine work, it is time for humans to lean into the human parts.

Public discourse about AI often swings between hyperbolic enthusiasm ("disrupting everything") and categorical skepticism ("full of hallucinations"). In light of our findings for introductory physics instruction, we must concede there is genuine disruption. With each new version, AI is cranking up the heat, but the good news is that none of this diminishes physics; it foregrounds what makes the discipline valuable: modeling the world, arguing from evidence, and making principled approximations. With a deliberate jump now, first-year courses can become *more* authentic, humane, and rigorous—not despite AI but because we choose to teach and value what AI cannot yet substitute for.

Supplementary material

Readers can access the supplementary material at *TPT Online*.

References

1. G. Kortemeyer, "Could an artificial-intelligence agent pass an introductory physics course?" *Phys. Rev. Phys. Educ. Res.* **19**, 010132 (2023).
2. G. Kortemeyer, M. Babayeva, G. Polverini, R. Widenhorn, and B. Gregorcic, "Multilingual performance of a multimodal artifi-

cial intelligence system on multisubject physics concept inventories," *Phys. Rev. Phys. Educ. Res.* **21**, 020101 (2026).
3. OpenAI, "Introducing GPT-5," https://openai.com/index/introducing-gpt-5/, accessed Aug. 2026.
4. G. Kortemeyer, "The losing battle against plug-and-chug," *Phys. Teach.* **54**, 14–17 (2016).
5. G. Polverini and B. Gregorcic, "Performance of ChatGPT on the test of understanding graphs in kinematics," *Phys. Rev. Phys. Educ. Res.* **20**, 010109 (2024).
6. R. J. Beichner, "Testing student interpretation of kinematics graphs," *Am. J. Phys.* **62**, 750–762 (1994).
7. M. T. Chi, P. J. Feltovich, and R. Glaser, "Categorization and representation of physics problems by experts and novices," *Cognit. Sci.* **5**, 121–152 (1981).
8. S. F. Wolf, D. P. Dougherty, and G. Kortemeyer, "Empirical approach to interpreting card-sorting data," *Phys. Rev. Spec. Top. Phys. Educ. Res.* **8**, 010124 (2012).
9. S. F. Wolf, D. P. Dougherty, and G. Kortemeyer, "Rigging the deck: Selecting good problems for expert-novice card-sorting experiments," *Phys. Rev. Spec. Top. Phys. Educ. Res.* **8**, 020116 (2012).
10. G. Kortemeyer, E. Kashy, W. Benenson, and W. Bauer, "Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses," *Am. J. Phys.* **76**, 438–444 (2008).
11. W. K. Adams, K. K. Perkins, M. Dubson, N. D. Finkelstein, and C. E. Wieman, "The design and validation of the Colorado Learning Attitudes about Science Survey," *AIP Conf. Proc.* **790**, 45–48 (2005).
12. S. B. McKagan et al., "PhysPort use and growth: Supporting physics teaching with research-based resources since 2011," *Phys. Teach.* **58**, 465–469 (2020).
13. OpenAI, "Sora," https://sora.chatgpt.com/, accessed Aug. 2026.
14. G. Kortemeyer and W. Bauer, "Cheat sites and artificial intelligence usage in online introductory physics courses: What is the extent and what effect does it have on assessments?" *Phys. Rev. Phys. Educ. Res.* **20**, 010145 (2024).
15. M. Halaweh and G. El Refae, "Examining the accuracy of AI detection software tools in education," in *2024 Fifth International Conference on Intelligent Data Science Technologies and Applications* (IDSTA) (IEEE, 2024), pp. 186–190.
16. C. Wieman, "How to become a successful physicist," *Phys. Today* **75**, 46–52 (2022).
17. H. Lin, "Learning physics vs. passing courses," *Phys. Teach.* **20**, 151–157 (1982).
18. G. Kortemeyer, J. Nöhl, and D. Onishchuk, "Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study," *Phys. Rev. Phys. Educ. Res.* **20**, 020144 (2024).
19. G. Kortemeyer and J. Nöhl, "Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study," *Phys. Rev. Phys. Educ. Res.* **21**, 010136 (2026).
20. Z. Chen and T. Wan, "Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy," *Phys. Rev. Phys. Educ. Res.* **21**, 010126 (2026).
21. M. N. Dahlkemper, S. Z. Lahme, and P. Klein, "How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT," *Phys. Rev. Phys. Educ. Res.* **19**, 010142 (2023).
22. C. Megowan-Romanowicz, "Whiteboarding: A tool for moving classroom discourse from answer-making to sense-making," *Phys. Teach.* **54**, 83–86 (2016).
23. K. Foote, A. Knaub, C. Henderson, M. Dancy, and R. J. Beichner, "Enabling and challenging factors in institutional reform: The case of SCALE-UP," *Phys. Rev. Phys. Educ. Res.* **12**, 010103 (2016).
24. P. Vogt, P. Sander, S. Küchemann, and J. Kuhn, "iPhysicsLabs meets AI@TPT: Analysis of the acceleration of a car using ChatGPT," *Phys. Teach.* **63**, 390–391 (2026).
25. D. R. Dounas-Frazer, K. L. Van De Bogart, M. R. Stetzer, and H. Lewandowski, "Investigating the role of model-based reasoning while troubleshooting an electric circuit," *Phys. Rev. Phys. Educ. Res.* **12**, 010137 (2016).
26. A. W. Robinson, "Don't just stand there — teach Fermi problems!" *Phys. Educ.* **43**, 83 (2008).
27. D. E. Meltzer, "Relation between students' problem-solving performance and representational format," *Am. J. Phys.* **73**, 463–478 (2005).
28. R. Chabay and B. Sherwood, "Computational physics in the introductory calculus-based course," *Am. J. Phys.* **76**, 307–313 (2008).
29. R. H. Landau, "Resource letter CP-2: Computational physics," *Am. J. Phys.* **76**, 296–306 (2008).
30. M. D. Caballero and L. Merner, "Prevalence and nature of computational instruction in undergraduate physics programs across the United States," *Phys. Rev. Phys. Educ. Res.* **14**, 020129 (2018).
31. R. W. Chabay and B. A. Sherwood, *Matter and Interactions* (John Wiley & Sons, 2015).
32. T. O. B. Odden, E. Lockwood, and M. D. Caballero, "Physics computational literacy: An exploratory case study using computational essays," *Phys. Rev. Phys. Educ. Res.* **15**, 020152 (2019).
33. C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, "Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?" *Phys. Rev. Spec. Top. Phys. Educ. Res.* **8**, 020104 (2012).
34. E. F. Redish, "Changing student ways of knowing: What should our students learn in a physics class," in *Proceedings of World View on Physics Education* (2005), pp. 1–13.
35. G. Kortemeyer and C. Westfall, "History of physics: Outing the hidden curriculum?" *Am. J. Phys.* **77**, 875–881 (2009).
36. P. Zhang, L. Ding, and E. Mazur, "Peer instruction in introductory physics: A method to bring about positive changes in students' attitudes and beliefs," *Phys. Rev. Phys. Educ. Res.* **13**, 010104 (2017).
37. C. H. Crouch and E. Mazur, "Peer instruction: Ten years of experience and results," *Am. J. Phys.* **69**, 970–977 (2001).

Gerd Kortemeyer *is a member of the rectorate of ETH Zurich and an associate of the ETH AI Center. He is also an Associate Professor Emeritus at Michigan State University. He holds a PhD in physics from Michigan State University, where he taught for two decades. His research focusses on technology-enhanced learning of STEM disciplines; currently, he is advancing the research and development of AI-based tools and workflows for teaching, learning, and assessment. He loves photography as one of his hobbies, but would never boil a frog for the sake of a non-AI version of Fig. 2.*
kortemey@msu.edu

DOI: 10.1119/5.0296601