Editors' Suggestion

Challenges in using online assessments to measure student understanding in the era of generative AI

David P. Smith,* Paula R. L. Heron, and Peter S. Shaffer
Department of Physics, University of Washington, Seattle, Washington 98195, USA

(Received 30 April 2026; accepted 13 October 2026; published 19 November 2026)

Many courses now incorporate online methods for assessing student learning. Some instructors assign questions prior to lecture (e.g., prelecture quizzes) to obtain information that helps them tailor their lessons to the current state of student understanding. To assess curriculum efficacy, physics education researchers often use "pretests," inventories, etc. that provide baseline data that can be compared to results from post-tests. These assessments were initially given on paper but moved online in the early 2000s. At the University of Washington, we found that the impact of this change was minimal; the results of our pretests stayed essentially the same when given on paper or online and were consistent from quarter to quarter and instructor to instructor. However, we have begun to observe systematic changes in pretest performance that are consistent with students using artificial intelligence (AI) tools to answer pretests, even though they are scored only on effort, not correctness. Using two-proportion z-tests, mixed-effects logistic regression, and Welch's t tests to analyze response length of explanations, this paper documents the nature and extent of these changes across 39 questions in two introductory algebra-based physics courses. This paper presents results that suggest caution to those using online quizzes to probe student understanding—both before and/or after instruction. We have seen increases in correct responses on pretests as great as 36%, but we have also seen significant decreases on questions that AI tools have tended to answer incorrectly. These findings raise important questions about the validity of online assessments and are intended to raise awareness for anyone using online tests to assess student learning. It remains an open question as to whether, and how, AI can be leveraged to improve instruction.

DOI: 10.1103/xrw4-t58t

I. INTRODUCTION

Studies of student learning often rely on comparing the state of student understanding before and after instruction. Assignments administered to students before instruction on a given topic identify the ideas and reasoning skills that students have at that stage in a course. The results can then be compared to those on a post-test given after instruction or an intervention to gauge the impact on student learning. The pre- and post-tests may be in the form of interviews, online or written questions, or a combination of each.

Since the early 1990s, the Physics Education Group at the University of Washington (UW PEG) has used pretests in introductory and advanced physics courses to probe student thinking before students have worked through materials developed by our group [1]. The results have demonstrated that some student ideas are particularly robust and do not change significantly before and after standard instruction unless certain incorrect or unproductive ideas are specifically targeted. The results are consistent from class to class, quarter to quarter, and instructor to instructor. This observation is supported by many other researchers using various tools, such as inventories, given to students across the country (e.g., the force concept inventory [2] or the conceptual survey of electricity and magnetism [3]). A paper by Heron [4] illustrates the historical consistency of pretest results at the UW, as well as the impact of lecture instruction on the results.

Since 2000, pretests at the UW have been administered online in the introductory calculus-based physics sequence. In 2016, they were extended to the algebra-based introductory sequence. In all cases, students are told they receive credit for effort, independent of whether their responses are correct. The pretests are often administered after lecture instruction on a given topic but prior to the corresponding tutorial. Since the public release of ChatGPT in November 2022, generative artificial intelligence (AI) tools have become increasingly available to students and have improved rapidly in capability. While pretest results remained relatively stable in the period immediately following that release, we began to observe notable changes around autumn 2024, consistent with

*Contact author: dsmith4@uw.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

growing student adoption of these tools. Since this would have a significant impact on how we and others interpret preinstruction assessments and evaluate curriculum, we began examining the pretest results across all our introductory courses to determine whether these changes were statistically reliable and whether their pattern was consistent with student use of AI tools.

Although there might be ways to leverage the use of AI on pretests to promote student learning, the intent of this study is to inform the physics education research (PER) community about both the nature and extent of these changes, even on low-stakes assignments. Specifically, we address the following research questions:

(1) Did the percentage of correct responses to pretest questions change significantly after autumn 2024, and if so, by how much?
(2) Are the observed changes in pretest performance, in both response accuracy and explanation length, consistent with student use of AI tools?

II. PRIOR RESEARCH

Preinstruction assessments have been used extensively for many decades to identify student learning difficulties with specific physics concepts, including in our prior work [5–10]. This method of assessing student understanding prior to instruction has been repeated by researchers throughout the United States and in many other nations [11–15]. Additionally, the use of concept inventories, such as those aforementioned, is widely used throughout the United States, both at the college and high-school level [16,17]. The cited studies represent only a very small subset of the extensive literature on preinstruction quizzes, and their relevance to physics education research is well established. We note that the conclusions made in the cited studies have relied on the assumption that student responses to the administered assessment tools reflect their own reasoning and prior knowledge.

Technological advances, including the widespread adoption of the Internet, have influenced both content delivery and classroom structure. These changes include the introduction of online homework systems such as MyLab and Mastering [18] and WebAssign [19], as well as publicly available simulations such as those provided by the PhET project [20]. Much research has examined how these tools affect student learning [21–24]. However, despite changes in delivery and access, these systems largely preserve the assumption that student responses reflect students' own thinking and understanding.

Many of the recent publications in PER journals have begun to examine the role of AI in teaching and learning. Some studies have considered the role of AI in grading [25,26], while others have considered the possibility of using it as a collaborative tool [27,28]. Particularly relevant to this work is the rapid development of large language models and their increasing capability in solving physics problems. For example, Kortemeyer and Bauer [29] report that, as of 2024, when students are solving problems, they predominantly rely on traditional Internet resources, and only a quarter of their survey respondents claimed to have utilized some minimal form of AI tools in the solution of their homework problems. In contrast, more recent work by Tschisgale *et al*. [30] demonstrates that systems such as ChatGPT can achieve performance exceeding that of human participants on Physics Olympiad problems. These results suggest both a rapid increase in AI capability and a corresponding increase in student use, a trend consistent with the changes in pretest performance we document in this paper.

In summary, a substantial body of prior research has established preinstruction assessments as tools for examining student understanding, whether delivered online or in person. In these contexts, student responses are typically interpreted as providing an authentic picture of student understanding. The growing availability of AI tools complicates this interpretation, making it unclear whether responses to preinstruction quizzes can still be viewed as measures of prior knowledge and reasoning.

III. CONTEXT AND METHODS

Pretests at the University of Washington are given almost every week in our introductory algebra- and calculus-based physics courses. They are associated with *Tutorials in Introductory Physics* [31], the materials our group has been developing to improve student learning in introductory physics courses. The pretests have multiple goals: to help students recognize what they do and do not understand before participating in the upcoming tutorial, to provide instructors and TAs with insight into the concepts that students find difficult, and to give baseline information about student understanding that can be used to assess the impact of the tutorials.

Our algebra- and calculus-based sequences each consist of three courses that are each taught every quarter: *mechanics, electricity and magnetism*, and *waves and optics.* In any quarter, there are between one and three lecture sections for each course. These are typically taught by different instructors. The enrollments in each section range from about 100 to 250 students. The algebra-based sequence is primarily for life science majors; the calculus-based sequence is mostly for preengineering and other science, technology, engineering, and mathematics majors. There is a set syllabus and sequence of topics for each course, in part to ensure that all students see the same material as they progress through the sequence. Faculty rotate through the courses and most share course materials (slides, clicker questions, suggested demonstrations, etc.).

The algebra-based sequence has been the more stable and uniform of the two sequences following the return to in-person instruction at the UW after the pandemic. In that sequence, the textbook and sequencing of *tutorials* have

not changed substantially. There has been greater variation in the calculus-based sequence, mostly as a result of a change in textbook that necessitated reordering some of the topics and revisions to some *tutorials*. Thus, the results presented here are from the first two courses in the algebra-based sequence. However, we have seen similar results in all our introductory courses.

There are 10 pretests in the algebra-based sequence that are analyzed in this study, since they had been given unchanged over an extended period. In total, we examined 39 questions across the first two courses in the sequence, *mechanics* (Phys 114) and *electricity and magnetism* (Phys 115). Our Phys 115 course also includes material on *thermodynamics*. Visual inspection of the longitudinal data indicated that the most pronounced changes in student performance began around autumn 2024, although in some cases, there might have been some signal prior to that quarter. While generative AI tools became publicly available in November 2022, early versions performed poorly on physics assessments. For example, ChatGPT scored 60% on the force concept inventory, corresponding to the entry threshold for Newtonian physics [32]. The release of more capable versions and education-specific tools through 2023 and 2024 is consistent with the pattern of performance changes we observe in our data. We therefore use autumn 2024 as our dividing point, the rationale for which is discussed further in Sec. IV B. Throughout this paper, we refer to quarters preautumn 2024 as "pre-AI" quarters and quarters postautumn 2024 as "post-AI" quarters, noting that these labels reflect our interpretation of the data rather than a confirmed causal claim.

Originally, we only examined the change in the percentage of correct responses to each question, but we noticed changes in student explanations as well. Some recent responses seemed longer and more formal, often structured in ways we had not seen before. We therefore also examined how the average length of student explanations was changing.

Finally, we asked ChatGPT (version 4.0) [33] (and similar programs) to answer many of the pretest questions. In most cases, it gave correct answers, but some were incorrect, and in these cases, we found a decrease in the percentage of correct student responses. Note, however, that even within the few quarters covered by this study, and with the introduction of ChatGPT-5 in August 2025, AI has progressed and some questions that AI tools initially answered incorrectly are now answered correctly. It was thus not always possible to determine if, on a particular pretest question, students would have obtained a correct or incorrect answer from the particular AI tool they used.

We first present an analysis of the impact on the percentage of correct responses in Sec. IV, which is followed by an analysis of the changes in explanation length in Sec. V.

IV. IMPACT ON PERCENTAGE OF CORRECT RESPONSES

Our awareness of the changes in pretest results began when a member of our group was conducting research on a *scaling* tutorial that is being developed for the algebra-based sequence. One question on the associated pretest probes student thinking about the effect of isometric scaling on an object's mass. [See Figs. 1(a) and 1(b).] For several years, the results from this pretest had been quite stable and independent of lecture instruction. However, starting in 2024, the percentage of correct responses increased dramatically. [See Fig. 1(c).] Prior to then, the percentage of correct answers ranged from about 20% to 40%. Correct responses increased in winter 2024 to 56% and again in autumn 2024 to 74%. This question was among those showing an earlier signal than the other pretests examined. We speculate that this may be because it is text-based, making it straightforward to copy and paste directly into an AI tool, whereas most other pretest questions include diagrams.

The question shown in Fig. 1(a) was repeated on a midterm exam in spring 2025, after students had completed the *scaling* tutorial and received lectures on this material, with 72% of students answering correctly. The fact that pretest performance rivaled post-test performance motivated us to examine the results from many of the pretests administered since the UW returned to in-class instruction following the COVID pandemic (autumn 2022 to the present).

As stated earlier, we examined 39 pretest questions from our algebra-based sequence, and the analysis highlighted that for almost all questions, the percentage of correct responses was greater for post-AI quarters. To evaluate the difference in correct responses, we conducted two complementary analyses.

First, we used a two-proportion z-test to determine whether differences between pre- and post-AI quarters were statistically significant. Effect sizes were estimated using Cohen's h values. These results are presented in Sec. IV A.

Second, to account for potential quarter-to-quarter variation before and after autumn 2024, we used a mixed-effects logistic regression model. In nearly all cases, the conclusions were consistent with those from the two-proportion z-test. These results are discussed in Sec. IV B.

A. Results from two-proportion z-test

To implement the two-proportion z-test, we pooled the number of correct responses across all pre-AI quarters to form a single pre-AI sample and repeated this procedure for the post-AI quarters. Each proportion therefore reflects the total number of correct responses divided by the total number of students in the corresponding period. The test

(a) Pretest question

A dwarf siren salamander grows in such a way that it increases its size equally in all dimensions. An adult salamander is three times larger in every dimension than an infant salamander.

What is the ratio of the adult's mass to the infant's mass?

(b) Sample student explanation for correct answer

"Since the adult salamander is 3 times larger in every dimension than an infant salamander that would mean the volume of the adult salamander would be 27 (3 x 3 x 3) times greater than the infant salamander. That means the ratio of the adult's mass to the infant's mass is 27:1."

(c) Pretest results over multiple years

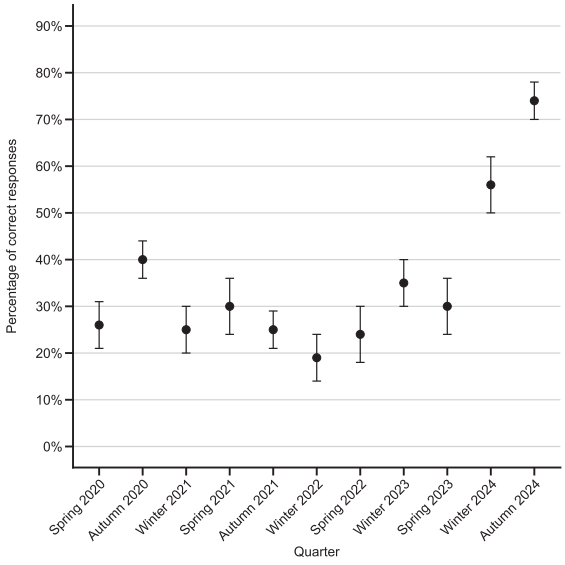
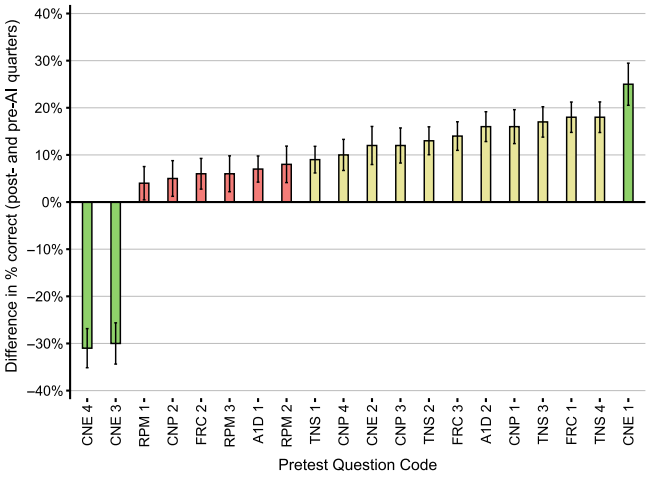


FIG. 1. (a) Pretest question for the *scaling* tutorial administered in the algebra-based mechanics course at the UW. (b) Sample of a correct student explanation. (c) Percentage of correct responses from spring 2020 to autumn 2024. A different version of the pretest was used in winter, spring, and autumn 2025, so those results are not shown. Each quarter includes responses from between 210 and 571 students, with an average of 330 students per quarter. Error bars show 95% confidence intervals using binomial standard error.

(a) Phys 114 Pretest Analysis



(b) Phys 115 Pretest Analysis

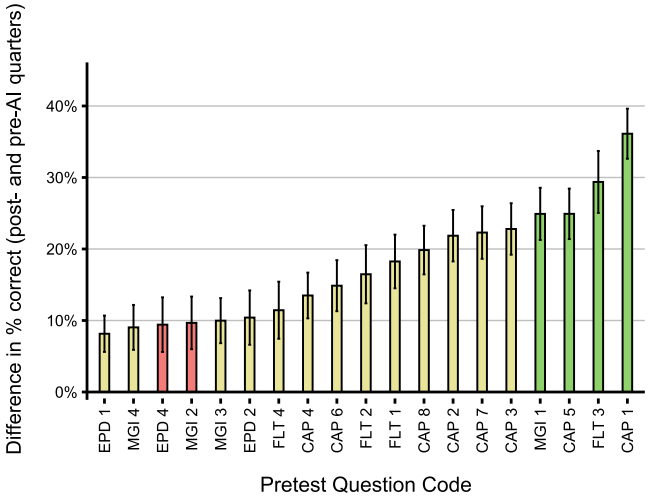


FIG. 2. Results from a two-proportion z-test assessing the change in the percentage of correct answers between pre- and post-AI quarters. The error bars show 95% confidence intervals for each question. Red indicates Cohen's h values less than 0.2, yellow indicates values between 0.2 and 0.5, and green shows values greater than 0.5. (a) Questions from the algebra-based course on *mechanics* (Phys 114). (b) Questions from the algebra-based course on *electricity and magnetism* (Phys 115).

evaluates whether the difference between these pooled proportions is statistically significant. Two-tailed p values using a Bonferroni-corrected significance threshold of $p < 0.05/20 = 0.0025$ for Phys 114 and $p < 0.05/19 \approx 0.0026$ for Phys 115.

In addition to statistical significance, we calculated Cohen's h to quantify effect size. This metric expresses how large the differences between two proportions are on a standardized scale. Following common guidelines, h values approximately equal to 0.2, 0.5, and 0.8 represent small, medium, and large effects, respectively.

Figs. 2(a) and 2(b) show the analysis for the pretest questions administered in our algebra-based *mechanics* course (Phys 114) and in our *electricity and magnetism* course (Phys 115), respectively. Table I provides a list of the examined pretests and the corresponding three-letter code used in Figs. 2(a) and 2(b).

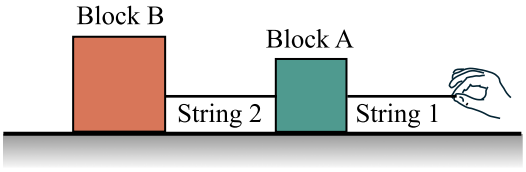
The results are presented in order of increasing change in the percentage of correct responses. After applying the

TABLE I. Three-letter codes for pretests used in this study.

Code	Pretest full name
A1D	Acceleration in one dimension
CAP	Capacitance
CNE	Conservation of energy
CNP	Conservation of momentum
EPD	Electric potential difference
FLT	First law of thermodynamics
FRC	Forces
MGI	Magnetic interactions
RPM	Representations of motion
TNS	Tension

Bonferroni correction, 18 out of 20 questions in Phys 114 and all 19 questions in Phys 115 showed statistically significant differences. The two questions in Phys 114 that did not meet the corrected significance threshold were those with the smallest Cohen's h values. Of the questions examined, 14 out of 20 questions in Phys 114 and 17 out of the 19 questions in Phys 115 had Cohen's h values corresponding to medium or large effects.

Two blocks of unequal mass ($m_A < m_B$) on a level, frictionless table are connected by a string (String 2). Another string (String 1) is attached to Block A. A hand pulls on String 1 with constant force such that the blocks move together to the right with increasing speed. Assume that both strings are inextensible and massless.



Q1 Is the magnitude of the acceleration of block A *greater than, less than,* or ***equal to*** the magnitude of the acceleration of block B?

Q2 Is the magnitude of the force exerted on block A by string 1 ($T_{1\text{ on }A}$) *greater than, less than,* or ***equal to*** the magnitude of the force exerted on string 1 by block A ($T_{A\text{ on }1}$)?

Q3 Is the magnitude of the force exerted on block A by string 2 ($T_{2\text{ on }A}$) *greater than, less than,* or ***equal to*** the magnitude of the force exerted on block B by string 2 ($T_{2\text{ on }B}$)?

Q4 Is the magnitude of the force exerted on block A by string 1 ($T_{1\text{ on }A}$) ***greater than***, *less than,* or *equal to* the magnitude of the force exerted on block B by string 2 ($T_{2\text{ on }B}$)?

FIG. 3. The four questions on the *tension* pretest. The correct answers are shown in green.

In Sec. IV A 1, we present a detailed analysis of a typical case. This is followed in Sec. IV A 2 by an analysis of the two questions in Phys 114, CNE 3 and CNE 4, which had large decreases in the percentage of correct responses. Excluding these questions, the average change in correct responses in Phys 114 was 12% and 18% in Phys 115.

1. Typical case

Questions 1 through 4 on the *tension* pretest in the algebra-based sequence (see Fig. 3) are typical of the questions for which performance changed by a moderate amount (10%–15%). The difficulties that students encounter when reasoning about the situation given in the pretest and the impact of the tutorials on student performance are discussed in an earlier paper [34].

Figure 4 shows the percentage correct on each question for each quarter from autumn 2022 through autumn 2025.

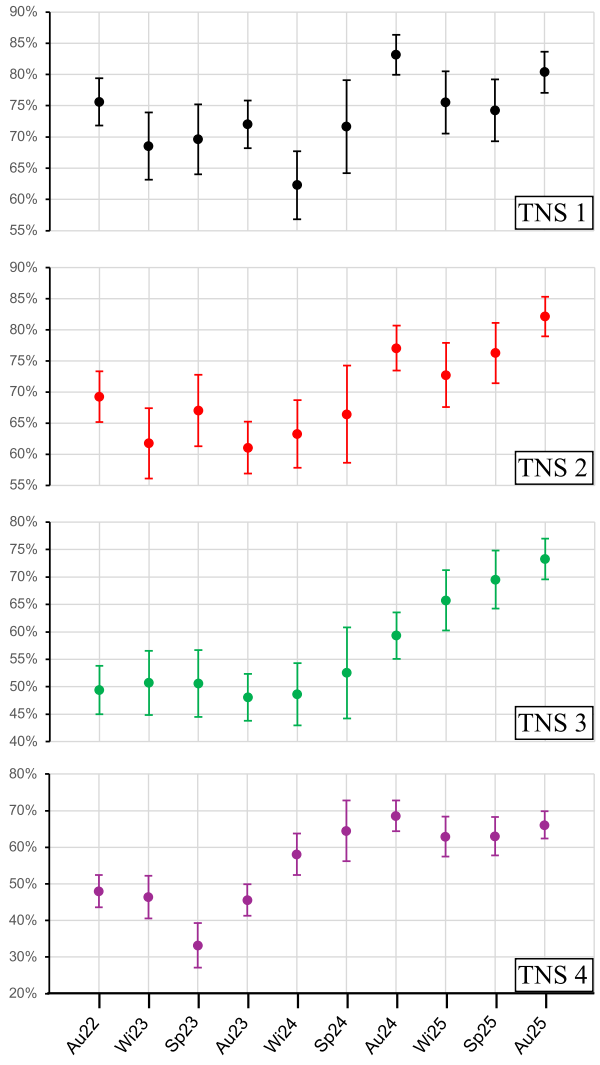
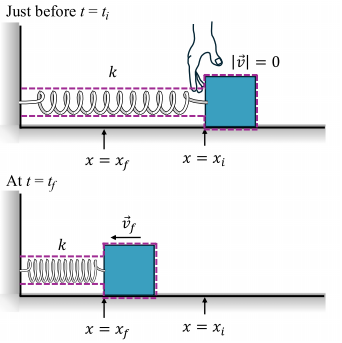


FIG. 4. Results by quarter for the questions in the *tension* pretest given in algebra-based sequence (Phys 114). Error bars show 95% confidence intervals using binomial standard error.

A block of mass m is attached to an ideal spring with spring constant k, as shown. The friction between the block and the surface is negligible.



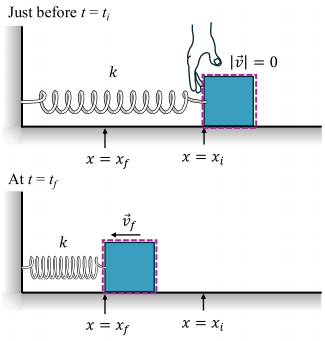
Initially, the block is at rest at position $x = x_i$, and the spring is stretched.

At time $t = t_i$, the block is released from rest. The block speeds up until $t = t_f$, when it passes x_f with speed v_f. In answering the following questions, consider system 1, consisting of the block and spring, as indicated by the dashed line shown.

Q1 During the interval from t_i to t_f, is the net work on system 1 by external forces positive, negative, or zero?

Q2 During the interval from t_i to t_f, does the total energy of system 1 increase, decrease or remain the same?

Now consider system 2, consisting of the block only, as indicated by the dashed line in the diagram. Recall that initially, the block is at rest at position $x = x_i$, and the spring is stretched. At time $t = t_i$, the block is released from rest. The block speeds up until $t = t_f$, when it passes x_f with speed v_f.



Q3 During the interval from t_i to t_f, is the net work on system 2 by external forces positive, negative or zero?

Q4 During the interval from t_i to t_f, does the total energy of system 2 increase, decrease, or remain the same?

FIG. 5. Pretest questions for the *conservation of energy* tutorial administered in the algebra-based mechanics course at the UW.

The increase in the percentage of correct responses is between 10% and 15% for the various questions starting in about autumn 2024.

ChatGPT produced correct answers to the *tension* pretest questions when the question text was pasted directly into the model. In the explanation for question 3, however, early versions of ChatGPT stated that the force exerted on block A by string 2 and the force exerted on block B by string 2 are third-law force pairs. Rather, they are equal in magnitude since the string is massless. In later versions, the explanation given by ChatGPT was correct. The percentage of students giving Newton's third law as an explanation for that question increased from 8% prior to autumn 2024 to 18% from autumn 2024 onward.

To determine the net work on system 2 (the block only) by external forces during the interval from t_i to t_f, let's analyze the forces acting on the block and the energy transfers involved.

Forces Acting on System 2

1. Spring Force: As the block is attached to the spring, during the interval from t_i to t_f, the spring exerts a restoring force that accelerates the block towards the equilibrium position.

2. External Forces: Since we are considering only the block as the system, we should analyze any forces acting on it externally. For this ideal scenario, we assume:
   - No applied external forces (other than the force exerted by the spring).
   - Negligible friction or air resistance.

Analysis of Work Done

- The work done by the spring force is internal to system 2, as it is a force exerted by the system (the spring attached to the block).
- There are no external forces doing work on the block since we have defined system 2 as the block alone.

Net Work on System 2

Since no external forces are acting on the block, **the net work done on system 2 by external forces during the interval from t_i to t_f is zero.**

The change in energy of the block is entirely due to the conversion of potential energy from the spring into kinetic energy, but this does not involve external work done on the block itself.

FIG. 6. Response generated by ChatGPT when asked to provide an answer to question 3 of the *conservation of energy* pretest. The incorrect conclusion provided is in red.

2. Case in which ChatGPT gave an incorrect answer

As we analyzed students' pretest responses, we noticed some cases in which the percentage of correct answers decreased starting in autumn 2024. The results for questions 3 and 4 in the *conservation of energy* pretest (CNE) in the Phys 114 algebra-based mechanics course are two such cases. To provide context for questions 3 and 4, questions 1 and 2 are also shown (see Fig. 5).

As shown in Fig. 2(a), we observed increases in student performance for questions 1 and 2 (CNE 1 and CNE 2), with question 1 showing the largest increase in correct

To analyze how the total energy of system 2 (the block only) behaves during the interval from t_i to t_f, we must consider the energy transformations that occur as the block moves under the influence of the spring force.

Components of Energy in System 2

1. **Kinetic Energy (KE):** The block gains kinetic energy as it accelerates towards the equilibrium position. The kinetic energy is given by:

$$\mathrm{KE} = \frac{1}{2}mv^2$$

where m is the mass of the block and v is its velocity.

2. **Potential Energy (PE)**: Initially, when the block is at position x_i, it has potential energy due to the spring being stretched:

$$\mathrm{PE} = \frac{1}{2}kx_i^2$$

As the block moves towards the equilibrium position, the potential energy stored in the spring decreases.

Energy Changes During the Interval

- **Initial State (t_i):** At the start of the interval, the block is at rest, so:
  - KE = 0 (since it is at rest)
  - PE is at its maximum due to the stretch in the spring.
- **Final State (t_f):** By the time the block reaches position (x_f) with speed (v):
  - KE has increased as the block speeds up.
  - PE has decreased as the spring returns to its equilibrium position.

Conservation of Energy

In this idealized scenario (with no external work and negligible friction), the **total mechanical energy** of the block does not change. It is conserved:

$$Total\ Energy = \mathrm{KE} + \mathrm{PE}$$

As the potential energy decreases due to the spring force acting on the block, the kinetic energy increases correspondingly.

Conclusion

During the interval from (t_i) to (t_f), the **total energy of system 2 remains the same**.

This is because the decrease in potential energy from the spring is exactly equal to the increase in kinetic energy of the block, ensuring that the total mechanical energy is conserved.

FIG. 7. Response generated by ChatGPT when asked to provide an answer to question 4 of the *conservation of energy* pretest. The incorrect conclusion provided is in red.

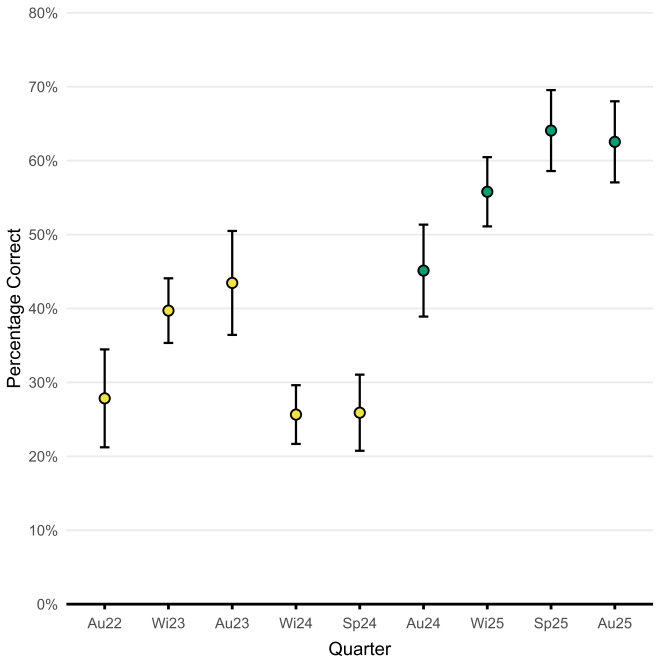


FIG. 8. The data above correspond to student responses to question 1 on the *magnetic interactions* tutorial pretest in Phys 115. The yellow markers correspond to the pre-AI cohort, and the green markers correspond to the post-AI cohort. Error bars show 95% confidence intervals using binomial standard error.

responses across all questions analyzed in Phys 114. For these questions, ChatGPT provided the correct answer, where both the net work done on system 1 and the change in total energy of the system are zero.

However, when we asked ChatGPT to answer questions 3 and 4, it provided the incorrect responses shown in Figs. 6 and 7, respectively.

B. Results from mixed-effects logistic regression

The two-proportion z-test confirmed that the differences we observed in the percentage of correct answers were statistically significant after applying the Bonferroni correction. Although this test provides an initial indication of a difference, we were concerned that this approach might be too coarse because it ignores the quarter-to-quarter variation that was visible in our data. An example of these fluctuations is shown in Fig. 8.

Because of these fluctuations, we sought a statistical method that could incorporate the fluctuations in the data more appropriately. We therefore used a mixed-effects logistic regression model. The dependent variable was the binary response (correct or incorrect) for each student on each question, and the pre- and postautumn 2024 periods were included as fixed effects. This model accounts for the differing sample sizes across quarters and includes a

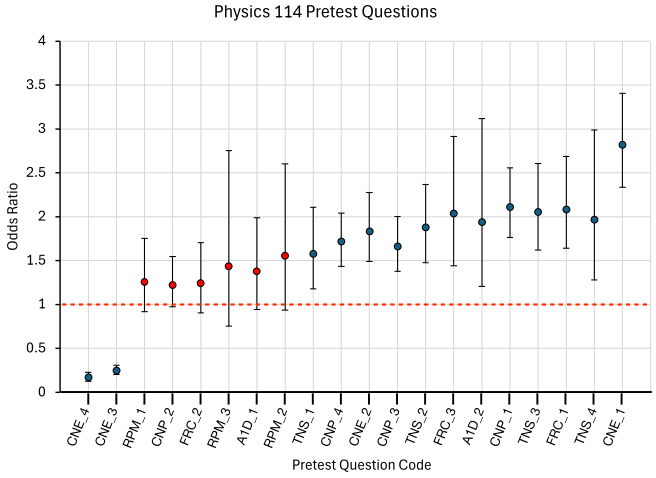


FIG. 9. Results from a mixed-effects logistic regression test assessing the change in the percentage of correct answers between pre- and post-AI quarters for pretests administered in Phys 114 (*mechanics*). Markers in red show odds ratio results that are not statistically significant, since the error bar includes the odds ratio of 1.

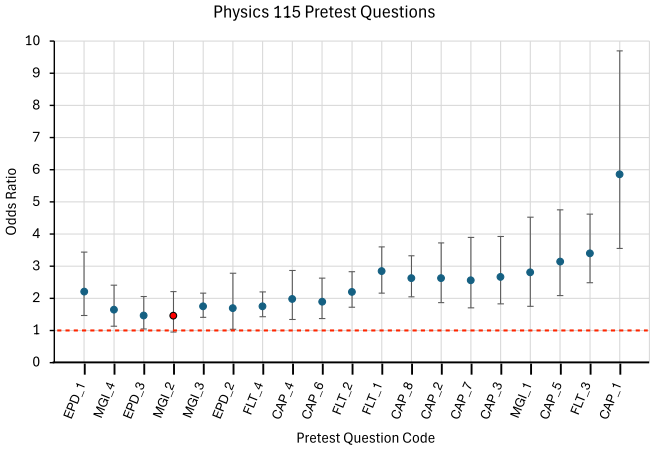


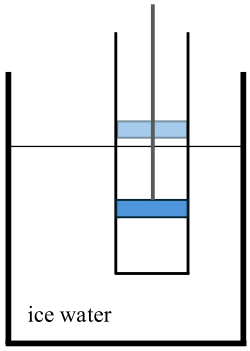
FIG. 10. Results from a mixed-effects logistic regression test assessing the change in the percentage of correct answers between pre- and post-AI quarters for pretests administered in Phys 115 (*electricity and magnetism*). Markers in red show odds ratio results that are not statistically significant.

random intercept by quarter, which allows the model to capture quarter-specific variation. No additional covariates were included. The output of this analysis is an odds ratio, which estimates how much more or less likely students in the post-AI cohort were to answer the question correctly compared to students in the pre-AI cohort. This provides a more detailed analysis than the two-proportion z-test, since it accounts for both the binary nature of the responses and the quarter-level variability.

In Fig. 2(a), we presented data on the difference in percentage correct for tutorial pretests administered in Phys 114 (*mechanics*). In that figure, the data were

Question 3 from *First Law of Thermodynamics* Tutorial Pretest

A sealed cylindrical pump contains one mole of an ideal gas. The piston fits tightly so that no gas escapes, but friction is negligible between the piston and the cylinder walls. The pump is placed in a very large container that is filled with an ice-water mixture.



The system is allowed to come to thermal equilibrium at 0°C. The piston is then pressed inward very slowly.

Will the temperature of the gas increase, decrease, or remain the same as the piston is pressed inward?

Student response

The temperature of the gas will remain the same as the piston is pressed inward because the system is in thermal equilibrium with a very large ice-water mixture at 0 degree Celsius. This ensures that the gas can exchange heat freely with its surroundings to maintain a constant temperature. As the piston is pressed inward very slowly, the compression work done on the gas causes heat to flow out of the gas into the ice-water mixture, balancing the energy changes. Since the process occurs slowly and the surrounding temperature remains constant, the process is isothermal, meaning the temperature of the gas does not change even though its volume and pressure are altered.

FIG. 11. Question 3 from the first law of thermodynamics pretest and a sample student response from a post-AI quarter.

organized from smallest to largest difference in percentage of correct answers. For clarity and to show the consistent results between both tests, the data in Fig. 9 below are organized in the same question order as Fig 1(a).

Although the two-proportion z-test resulted in p values indicating statistically significant differences for 18 of the 20 questions in Phys 114 after Bonferroni correction, the mixed-effects logistic regression test shows that 6 of the 20 questions have odds ratio values greater than 1, indicating an increase, but the 95% confidence interval includes an odds ratio of 1. In these cases, the model does not detect a statistically reliable difference between pre- and post-AI cohorts. However, the fact that 14 of the 20 questions had statistically reliable differences confirms that student performance in pre-AI quarters is different than that for post-AI quarters.

We carried out the same analysis on the pretest questions associated with Phys 115, and the results are shown in Fig. 10.

For the set of questions in Phys 115, only one question (MGI 2) had a 95% confidence interval range that included an odds ratio of 1.

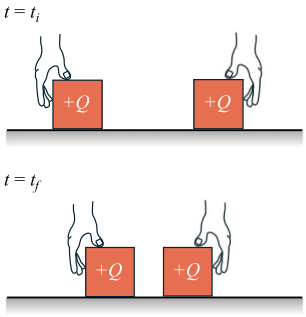
A broader discussion of the results presented in this section is provided in Sec. VI.

V. IMPACT ON EXPLANATION LENGTH

While examining changes in the percentage of correct responses, we also considered the explanations provided by students. In several instances, we noticed an increase in the formality of responses and the use of formatted equations. Two examples are shown in Figs. 11 and 12.

Question 1 on the *Electric Potential Difference* tutorial pretest

Two identical positively charged blocks are at rest on a frictionless surface. Two hands push the blocks toward each other from $t = t_i$ to $t = t_f$. The blocks begin and end at rest.



Does the electric potential energy of the two-block system above *increase, decrease,* or *remain the same* as a result of the movement?

Student explanation for their answer (increase) to question 1.

$F_{\text{electric}} = \frac{k|q_1||q_2|}{r}$, thus $F_{\text{electric}} \propto r$. The boxes carry identical, positive charges that remain the same from the start to the end of the scenario. However, the distance between them is decreased. Thus F_electric from one charge onto the other increases (magnitude) and is positive.

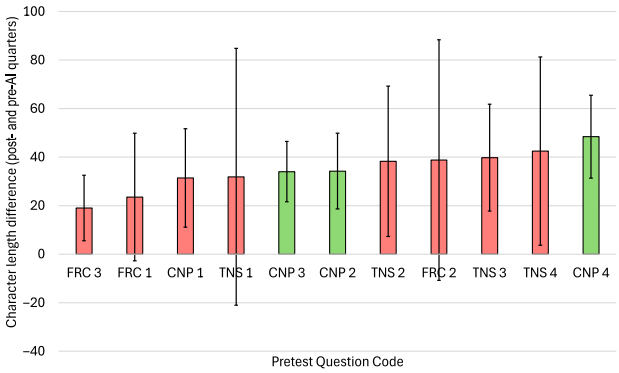
$W = F_{\text{electric}} \cdot d = \Delta U_{\text{electric}}$ given that the only force exerted on the system is electrical. Both blocks experience a positive electrical force and travel a certain distance. However, that movement was done by an external force exerted on each block. Thus, net positive external work is done onto the system. For the system to be at rest by the end of the scenario, meaning that K = 0 and the external work did not lead to a gain in kinetic energy, the system must have gained electric potential energy.

FIG. 12. Question 1 of the electric potential difference pretest and a sample student response from a post-AI quarter.

Although a full qualitative analysis of student explanations was beyond the scope of this study, the changes in formality shown in Figs. 11 and 12 motivated us to investigate whether explanation length also changed systematically before and after autumn 2024. We acknowledge that length alone cannot distinguish AI-generated from student-generated responses.

Analogous to using a two-proportion z-test to analyze changes in the correct response rate, we decided to use a Welch's t test to examine the statistical significance of changes in the length of explanations. For the Welch's t test, an average character length in students' explanations is determined for quarters prior to autumn 2024. The same calculation is done for quarters after autumn 2024, and these averages are then compared to determine whether the difference is statistically meaningful. We decided to

(a) Phys. 114: Change in character length of explanations corresponding to correct answers



(b) Phys. 114: Change in character length of explanations corresponding to incorrect answers

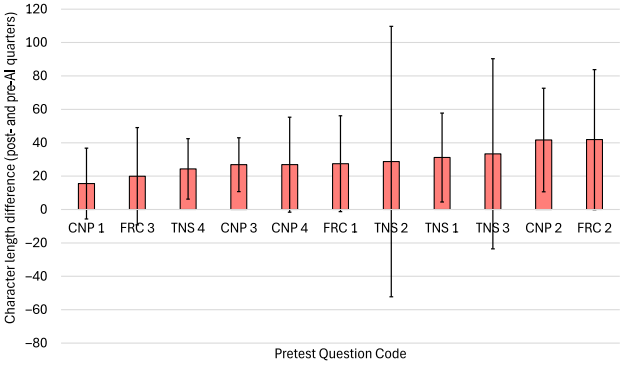


FIG. 13. Results from Welch's t test assessing the change in character length of explanations between pre- and post-AI quarters for cases in which the student provided the correct answer to the corresponding question. The error bars show 95% confidence intervals for each question. Red bars indicate explanation length differences that were not significant, and green bars indicate those questions that were significant. Error bars show 95% confidence intervals using binomial standard error.

separate explanations based on whether they were associated with a correct or incorrect answer to the corresponding multiple-choice question. As with the z-test, a Bonferroni correction was applied to account for multiple comparisons, adjusting the significance threshold to $p < 0.05/11 \approx 0.0045$ for Phys 114 and $p < 0.05/10 \approx 0.005$ for Phys 115.

The graphs in Figs. 13(a) and 13(b) show the analysis for Phys 114 pretests. We note that not all multiple-choice questions assessed for correctness had a follow-up question that asked students to explain their reasoning.

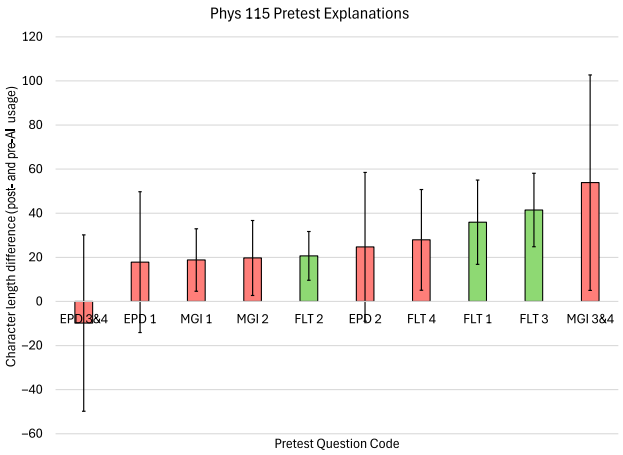
For the questions in Phys 114 that asked for explanations, we find that 3 of the 11 questions for which students gave a corresponding correct answer showed significant changes in explanation length after Bonferroni correction. For this set of 11 questions, the average change was 35 characters. None of the questions for which students gave an incorrect answer showed significant changes after Bonferroni correction.

We carried out the same analysis for questions in Phys 115, and the results are shown in Fig. 14.

For the 10 questions that asked for explanations in Phys 115, and for which students provided a correct answer to the previous multiple-choice question, 3 showed a statistically significant change in length after Bonferroni correction.

While the changes in explanation length alone do not provide strong evidence of AI use, we note that the changes in explanation length for correct answers are typically larger than those for incorrect answers, even where neither reaches significance. This pattern, together with the qualitative observations in Figs. 11 and 12 and the broader changes in correct response rates reported in Sec. IV, is consistent with the possibility that students are using AI tools to complete pretest questions.

(a) Phys. 115: Change in character length of explanations corresponding to correct answers



(b) Phys. 115: Change in character length of explanations corresponding to incorrect answers

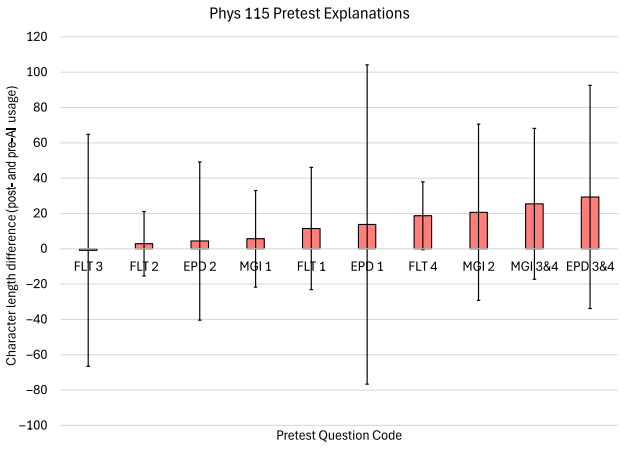


FIG. 14. Results from Welch's t test assessing the change in character length of explanations between pre- and post-AI quarters. The error bars show 95% confidence intervals for each question. Red bars indicate explanation length differences that were not significant, and green bars indicate those questions that were significant.

VI. DISCUSSION

Assessments of student learning play a critical role in making systematic improvements to instruction. Such assessments often rely on comparisons of results between pretests and post-tests. The feedback obtained not only gives a sense of the impact of an instructional intervention or curriculum but ideally reveals details that can guide iterative improvements. The results from this research provide insight into the potential impact of AI tools on research that relies on online tools for collecting data. Online proctoring tools such as Honorlock [35] and Respondus Monitor [36] are available for high-stakes exams, but their use is generally impractical for low-stakes, effort-graded assessments such as pretests, both because the level of surveillance is disproportionate to the stakes and because students can circumvent browser-based controls using a second device.

The fact that we saw changes in student performance by as much as 36%, even on surveys that are not graded for correctness, is concerning. The average change across all questions was approximately 15%, and this effect may increase as AI tools become more widely used and more capable. For curriculum developers, gains of this size between pre- and post-tests would typically be interpreted as evidence of a meaningful instructional impact. However, our findings indicate that comparison of results from online pretests with those of written post-tests may instead hide or eliminate apparent gains in student understanding.

One possible response is to instruct students not to use AI when completing pretests, but such instructions are likely to be ineffective. For example, a graduate student in

our group recently administered a paper-based pretest in a large class and asked students not to use any Internet-connected devices, yet observed a student using their phone to photograph the assessment and uploading it to an AI application. Although anecdotal, this example highlights the challenges of obtaining authentic student responses to assessments.

While one could argue that student use of AI may, in some cases, support engagement with course material and contribute to learning, this does not resolve the issue of its use on assessments. As others have begun to explore, interaction with AI tools may provide a more collaborative or coachlike experience, and it raises the possibility of modifying instructional materials, such as tutorials, to incorporate this approach. As such, further work is needed to determine how AI tools can be productively integrated into instruction while also preserving the validity of assessment data.

Based on our findings, we suggest that the PER community consider the purpose of a given assessment when deciding how to administer it. When pretests or concept inventories are used for research purposes, such as establishing baseline data or evaluating curriculum efficacy, we recommend administering them in person and on paper, with appropriate proctoring, acknowledging that even this approach presents challenges as AI tools become more accessible. When assessments are used primarily for pedagogical purposes, such as priming students for an upcoming tutorial or encouraging self-reflection on prior knowledge, online administration may remain appropriate, since the primary value lies in the engagement rather than the accuracy of the data.

Our findings highlight the need for the PER community to carefully consider how preinstruction assessments are administered and how student performance on such assessments is interpreted.

ACKNOWLEDGMENTS

The authors would like to thank the members of the UW PEG for their feedback during the process of preparing this paper and Paul van Kampen for a thorough reading of the manuscript. This material was based upon work supported by the National Science Foundation under Grant No. DUE-1821032.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they contain sensitive personal information. The data are available from the authors upon reasonable request.

[1] L. C. McDermott, Oersted medal lecture 2001: "Physics education research—The key to student learning," Am. J. Phys. **69**, 1127 (2001).
[2] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).
[3] D. P. Maloney, T. L. O'Kuma, C. J. Hieggelke, and A. Van Heuvelen, Surveying students' conceptual knowledge of electricity and magnetism, Am. J. Phys. **69**, S12 (2001).
[4] P. R. L. Heron, Effect of lecture instruction on student performance on qualitative questions, Phys. Rev. ST Phys. Educ. Res. **11**, 010102 (2015).
[5] P. S. Shaffer and L. C. McDermott, A research-based approach to improving student understanding of the vector nature of kinematical concepts, Am. J. Phys. **73**, 921 (2005).
[6] K. Kellar and P. Heron, Distinguishing between students' conceptual understanding and reasoning approaches: An application of dual-process theories, Phys. Rev. Phys. Educ. Res. **21**, 010141 (2025).
[7] H. G. Close and P. R. L. Heron, Research as a guide for improving student learning: An example from momentum conservation, Am. J. Phys. **78**, 961 (2010).
[8] B. A. Lindsey, P. R. L. Heron, and P. S. Shaffer, Student ability to apply the concepts of work and energy to extended systems, Am. J. Phys. **77**, 999 (2009).
[9] B. S. Ambrose, P. R. L. Heron, S. Vokos, and L. C. McDermott, Student understanding of light as an electromagnetic wave: Relating the formalism to physical phenomena, Am. J. Phys. **67**, 891 (1999).
[10] R. E. Scherr, P. S. Shaffer, and S. Vokos, Student understanding of time in special relativity: Simultaneity and reference frames, Am. J. Phys. **69**, S24 (2001).
[11] C. Furió and J. Guisasola, Difficulties in learning the concept of electric field, Sci. Ed. **82**, 511 (1998).
[12] L. Bollen, P. van Kampen, C. Baily, M. Kelly, and M. De Cock, Student difficulties regarding symbolic and graphical representations of vector fields, Phys. Rev. Phys. Educ. Res. **13**, 020109 (2017).
[13] W. M. Christensen and J. R. Thompson, Investigating graphical representations of slope and derivative without a physics context, Phys. Rev. ST Phys. Educ. Res. **8**, 023101 (2012).
[14] Q. X. Ryan, B. R. Wilcox, and S. J. Pollock, Student difficulties with boundary conditions in the context of electromagnetic waves, Phys. Rev. Phys. Educ. Res. **14**, 020126 (2018).
[15] D. P. Smith and P. van Kampen, Teaching electric circuits with multiple batteries: A qualitative approach, Phys. Rev. ST Phys. Educ. Res. **7**, 020115 (2011).

[16] J. Von Korff, B. Archibeque, K. A. Gomez, T. Heckendorf, S. B. McKagan, E. C. Sayre, E. W. Schenk, C. Shepherd, and L. Sorell, Secondary analysis of teaching methods in introductory physics: A 50 k-student study, Am. J. Phys. **84**, 969 (2016).

[17] A. Madsen, S. B. McKagan, and E. C. Sayre, Resource letter RBAI-1: Research-based assessment instruments in physics and astronomy, Am. J. Phys. **85**, 245 (2017).

[18] Mastering digital learning platforms | Pearson, https://www.pearson.com/en-us/higher-education/products-services/mastering.html.

[19] WebAssign, https://www.webassign.net/index.html.

[20] PhET interactive simulations, https://phet.colorado.edu/.

[21] A. Titus and G. Tang, Increasing students' time on task in calculus and general physics courses through webassign, in *2002 Annual Conference Proceedings* (ASEE Conferences, Montreal, Canada, 2002), pp. 7.660.1–7.660.20.

[22] S. B. McKagan, K. K. Perkins, M. Dubson, C. Malley, S. Reid, R. LeMaster, and C. E. Wieman, Developing and researching PhET simulations for teaching quantum mechanics, Am. J. Phys. **76**, 406 (2008).

[23] C. E. Wieman, W. K. Adams, P. Loeblein, and K. K. Perkins, Teaching physics using PhET simulations, Phys. Teach. **48**, 225 (2010).

[24] K. Lazarova, The role of online homework in low-enrollment college introductory physics courses, J. Coll. Sci. Teach. **44**, 17 (2015).

[25] G. Kortemeyer, Toward AI grading of student problem solutions in introductory physics: A feasibility study, Phys. Rev. Phys. Educ. Res. **19**, 020163 (2023).

[26] R. Mok, F. Akhtar, L. Clare, C. Li, J. Ida, L. Ross, and M. Campanelli, Using large language models for grading in education: An applied test for physics, Phys. Educ. **60**, 035006 (2025).

[27] D. Tong, B. Jin, Y. Tao, H. Ren, A. Y. M. Atiquil Islam, and L. Bao, Exploring the role of human-AI collaboration in solving scientific problems, Phys. Rev. Phys. Educ. Res. **21**, 010149 (2025).

[28] S. Kilde-Westberg, A. Johansson, and J. Enger, Generative AI as a lab partner: A case study, Phys. Rev. Phys. Educ. Res. **21**, 020119 (2025).

[29] 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).

[30] P. Tschisgale, H. Maus, F. Kieser, B. Kroehs, S. Petersen, and P. Wulff, Evaluating GPT- and reasoning-based large language models on physics olympiad problems: Surpassing human performance and implications for educational assessment, Phys. Rev. Phys. Educ. Res. **21**, 020115 (2025).

[31] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics* (Pearson College Div, Upper Saddle River, New Jersey, 2002).

[32] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, Phys. Rev. Phys. Educ. Res. **19**, 010132 (2023).

[33] ChatGPT, https://chatgpt.com/.

[34] L. C. McDermott, P. S. Shaffer, and M. D. Somers, Research as a guide for teaching introductory mechanics: An illustration in the context of the Atwood's machine, Am. J. Phys. **62**, 46 (1994).

[35] Honorlock online proctoring services and software, https://honorlock.com/.

[36] Prevent cheating in higher Ed with respondus lockdown browser, https://web.respondus.com/he/lockdownbrowser/.