Investigating student ability to follow reasoning chains: The role of conceptual understanding

Beth A. Lindsey*
Penn State Greater Allegheny, McKeesport, Pennsylvania 15132, USA

MacKenzie R. Stetzer
University of Maine, Orono, Maine 04469, USA

J. Caleb Speirs
University of North Florida, Jacksonville, Florida 32224, USA

William N. Ferm, Jr.
Ellsworth High School, Ellsworth, Maine 04605, USA

Alexander van Hulten
Penn State Greater Allegheny, McKeesport, Pennsylvania 15132, USA

(Received 21 November 2022; accepted 27 February 2023; published 20 April 2023)

In this paper, we seek to evaluate the extent to which students can follow a deductive reasoning chain when it is presented to them. A great deal of instruction in introductory physics courses is centered on presenting students with a logical argument that starts from first principles and systematically leads to a particular conclusion. This approach to instruction may conflict with current models of how students reason, including dual-process theories of reasoning and decision making. We investigated student ability to follow reasoning chains at several different institutions, across multiple topics that span introductory mechanics and electricity and magnetism. (For the purposes of this study, we operationally define "following" a reasoning chain—either a correct chain or an incorrect chain—as selecting the appropriate conclusion to a given chain.) To accomplish this, we asked students to answer a physics question and provide an explanation for their answer. We then presented them with a reasoning chain generated by a fictitious student and asked them to select the appropriate concluding statement for that chain. Some of these fictitious chains were fully correct, while others contained a conceptual or logical error. We intentionally used tasks for which students would be unlikely to generate correct reasoning chains on their own. We found that for most tasks, students were generally successful at following chains based on common incorrect reasoning. Students who themselves generated the correct reasoning, however, were much more successful at following correct reasoning chains. Connections between this work and dual-process theories, as well as implications for instruction, are discussed.

DOI: 10.1103/PhysRevPhysEducRes.19.010128

I. INTRODUCTION

Physics instructors often present explanations in logical chains, beginning from first principles and progressing toward a conclusion. While this approach is sensible from the viewpoint of both teachers and students, evidence is increasingly mounting that students often do not actually reason via such chains when presented with a qualitative physics question [1–3]. Indeed, experienced instructors recognize that, in many situations, their students seem to adopt an "answer first, explanation later" approach when responding to physics questions. The differences between the content presented in a traditional physics lesson and the conceptual understanding of that content that students develop, and the need to "meet students where they are"—both conceptually, as well as in how they frame the task at hand—have been well documented [4–6]. The question thus naturally arises as to how the mismatch between student reasoning patterns and the typical logical explanations presented in class affects the value that students draw from these explanations.

Researchers have begun to draw on theories from cognitive science to develop new models that better account

*bal23@psu.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.

for documented student reasoning patterns. Dual-process theories of reasoning and decision making [7,8] have recently been used to predict and explain student performance on physics tasks [1,9], as well as to develop interventions to improve student performance on certain qualitative questions [10]. These theories model human cognition as consisting of two processes: a quick, intuitive process and a slow, effortful, analytical process. Dual-process theories posit that, when a student approaches a question, their intuition presents them with a first-available mental model suggesting a particular response. Their analytic process may or may not activate to evaluate this model. If the analytic process does not intervene, the response suggested by the first-available mental model becomes the final response. Even if the analytic process does intervene, it is subject to a number of reasoning biases and may not reject an inappropriate conclusion or generate a correct model.

Despite evidence that some students do not tend to reason themselves using deductive reasoning chains, a great deal of instruction in physics nevertheless commonly includes presentations of such chains. In fact, several popular research-based instructional materials for introductory physics [11–13] rely on dialogues between two or more fictitious students as an instructional tool, in which students are expected to read statements and assess the validity of the arguments (i.e., reasoning chains) presented. Such "student dialogues" typically present both correct and incorrect arguments, and the statements are frequently constructed from common student responses to the question at hand. The utility of this instructional tool is based on the premise that students will follow the reasoning chains given by the fictitious students. To date, however, very few studies have examined the extent to which students, when presented with a reasoning chain, can follow the chain in order to infer the appropriate conclusions from the given argument, particularly if this argument is in conflict with their own first-available mental model. As such, the goal of the present study is to provide insight into students' ability to follow reasoning chains, with the ultimate aim of leveraging the findings to further enhance research-based instructional strategies.

The primary research questions driving this work are the following:

1. When presented with a typical reasoning chain (either correct or incorrect), do students follow it, thereby inferring the appropriate conclusion?
2. How does students' conceptual understanding interact with their ability to follow a reasoning chain?

To answer these questions, we first presented students with a physics task and asked them to answer it for themselves and to explain their reasoning. We used their response and explanation as a proxy for their own conceptual understanding of the task. We then presented students with a reasoning chain generated by a fictitious student and asked them to select the appropriate conclusion to this chain. To mimic the student dialogues found in many research-based instructional materials, we included both correct chains and incorrect chains that contained a conceptual or logical error. For the purposes of this work, we operationally define following a reasoning chain as selecting the appropriate conclusion implied by the presented chain of reasoning. This might also be referred to as "mapping" the appropriate conclusion onto a given reasoning chain. We have previously shown that students can follow qualitative, inferential chains involving a few simple steps. [14] The present work extends this analysis to examine student ability to follow more complex qualitative inferential reasoning chains, with a particular focus on conceptually difficult tasks for which students are unlikely to independently generate a correct reasoning chain. Since both correct and incorrect reasoning chains are commonly used as instructional tools, we included reasoning chains that were based on flawed premises or logic in addition to correct and complete reasoning chains.

II. MOTIVATION AND THEORETICAL BACKGROUND

While many instructors may expect that students will begin from first principles and reason deductively through a logical chain in order to arrive at the solution to a question or problem, an emerging body of research suggests that student cognition can be better modeled using dual-process theories of reasoning and decision making [1,15,16]. In this section, we provide an overview of these theories. While the assessment approach described in this work can be understood without this background in dual-process theories, these theories provide important motivation for why this work has been undertaken.

A. Dual-process theories of reasoning and decision-making

Dual-process theories have emerged from the field of cognitive psychology. The theories suggest that two distinct cognitive processes are involved in reasoning and making decisions. The first process is known as "process 1" or the "heuristic process" and is characterized by being fast and intuitive. The other process is known as "process 2" or the "analytic process" and is slow and deliberate. Process 1 is automatic; it will generate a mental model without conscious effort whenever a student is presented with a physics question. The first-available model from process 1 will be heavily influenced by a student's prior knowledge and beliefs, and also by contextual cues. For instance, the presence or absence of a highly salient, irrelevant feature in the representation of the task at hand can influence the response generated by process 1 [17]. Process 2, on the other hand, is effortful [8]. To describe the interactions between the two processes, we draw on the

heuristic-analytic theory of reasoning put forward by Evans [15]. According to this theory, when presented with a question, the heuristic process will generate one mental model at a time (the *singularity principle*). The initial mental model will be informed by the reasoner's content knowledge, but also by specific features of the task, and their own expectations for the goal of the task at hand. The role of the analytic process is to evaluate this model by ascertaining whether or not it is satisfactory for the task at hand (the *satisficing principle*). The analytic process, however, is also subject to a number of reasoning biases, such as a confirmation bias and other reasoning shortcuts. It will tend to accept the response generated by the heuristic process unless it identifies a compelling reason to reject it. Thus, the response suggested by the first-available model generated by the heuristic process tends to be the default response.

In order to reject an incorrect first-available mental model, the analytic process must detect a need to override this model and then must sustain this override to facilitate the consideration and adoption of an alternate model [18]. In order to detect the need to override, a student must possess the relevant knowledge and skills to answer the question correctly. In an analogy to computer software, these relevant knowledge and skills are frequently referred to as "mindware" [19]. For example, in order to compare the kinetic energies of two carts of different mass that are pushed by the same force through the same distance, a student must have mindware that includes the definition of work, the ability to recognize the need to use the work-energy relation, and the ability to apply that relation. In the absence of this understanding, the student would have no reason to question any model generated by the heuristic process that is based on other features of the base task, such as comparing the masses of the carts as in the reasoning chain shown in Fig. 1(d). Even if the student possesses some level of the appropriate mindware, that student might not access that mindware in every appropriate situation. For instance, a student who has demonstrated the ability to calculate work correctly, and to equate that work to a change in kinetic energy in a previous problem, might not think to apply these ideas to the base task shown in Fig. 1(a). Many factors may impact whether or not a student thinks to apply the "work" mindware to the base task, including students' general mindset and their expectations for the task at hand [15] or the presence of any salient distracting features in the task [9,17].

In addition to different levels of mindware, individual students may also exhibit differences in their tendency to critically evaluate mental models, referred to as cognitive reflection [20]. A student may not recognize that the response generated from their initial mental model is inconsistent with the work mindware for the very same reasons that prevented them from generating the correct model in the first place. The cognitive reflection test (CRT) has been developed and employed to gauge the tendency toward such reflective thinking [20]. A student with a greater tendency toward cognitive reflection may be more likely than others to reject an incorrect first-available mental model in favor of a correct model, even in the presence of similar levels of mindware. Indeed, research has shown evidence for an association between scores on the CRT and on other assessments such as the Force Concept Inventory [21]. Similarly, research has shown that students with lower CRT scores are more likely to give an intuitively appealing incorrect answer to certain conceptual questions [22] and may derive less benefit from some research-based instructional interventions [23].

B. Motivation for the present study

As noted previously, a great deal of instruction in physics involves starting from first principles and reasoning deductively to reach a logical conclusion. This mode of instruction, however, is at odds with the actual mechanisms of human reasoning posited by dual-process theories. In a great many cases, students do not appear to naturally start from first principles and reason deductively to reach a conclusion. While their process 2 reasoning may allow them to generate such a solution in support of a model presented by process 1, the process of following a reasoning chain presented by an instructor (or other instructional materials) that responds to a particular question is different than the process of generating and critically evaluating a mental model that provides a response to that question. Even if, as dual-process theories suggest, students do not necessarily themselves reason by generating logical deductive chains, they may nevertheless be capable of processing and analyzing such a chain when it is presented to them. Investigating that question is the subject of the present work.

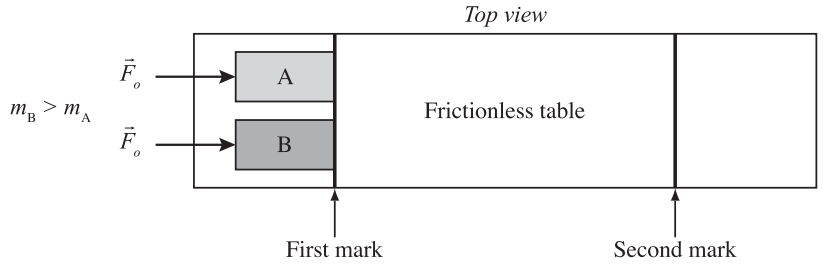
III. RESEARCH METHODS AND TASK DESIGN

A. The "follow" tasks: Overview of structure

The tasks used in this study all followed the same general structure. Each task was administered online and presented over multiple pages; students were not able to return to or edit any page of the task after it was completed. Students were first shown a qualitative reasoning task, or base task, and asked, in a multiple-choice format, what answer they themselves would give to this base task. (In each case, questions were posed such that the multiple-choice options given spanned the entire range of possible answers, for instance, by asking whether one quantity is greater than, less than, or equal to another or whether the comparison cannot be determined from the given information.) Students were also asked to explain their reasoning in an accompanying text box.

After completing the base task, students were told, "Now, another student assembles a line of reasoning to

(a)



Two carts, A and B, are initially at rest on a frictionless, horizontal table. A constant force of magnitude F_o is exerted on each cart as it travels from the first mark on the table to the second, after which each cart glides freely . The mass of cart A is less than that of cart B.

Is the kinetic energy of cart A greater than, less than, or equal to that of cart B after the carts have passed the second mark?

(b)

Now, another student attempts to answer the question. The student uses the following reasoning:

1. The same force is exerted on each cart.
2. The force acts through the same distance in each case.
3. Work is equal to the force dotted with the displacement: W = $\boldsymbol{F} \cdot \boldsymbol{d}$
4. Therefore, the same amount of work is done on each cart.
5. Work done on a system causes the energy of that system to increase.
6. In this case, the only type of energy changing for either cart is kinetic energy.

(c)

Now, another student attempts to answer the question. The student uses the following reasoning:

1. The same force is exerted on each cart.
2. The lighter cart will have a greater acceleration than the heavier cart.
3. By $v_f^2 = 2ad$, the cart with the greater acceleration will have a larger velocity when it crosses the second mark.
4. Kinetic energy is equal to $1/2\ mv^2$.
5. Velocity impacts kinetic energy more than mass because the velocity is squared to get the kinetic energy, but mass isn't.

(d)

Now, another student attempts to answer the question. The student uses the following reasoning:

1. Kinetic energy is equal to $1/2\ mv^2$.
2. Therefore, kinetic energy is proportional to mass.

FIG. 1. Follow task 1, the *work and kinetic energy* follow task. The base task is shown in (a). The reasoning chain shown in (b) is correct and supports the correct conclusion that the kinetic energy of cart A will be equal to cart B after the carts pass the second mark. The reasoning chain shown in (c), the *v-wins chain*, supports the conclusion that the kinetic energy of cart A will be greater than the kinetic energy of cart B. The reasoning chain shown in (d), the *mass-only chain*, supports the conclusion that the kinetic energy of cart A will be less than the kinetic energy of cart B.

answer the question." They were presented with the fictitious student's reasoning in the form of an enumerated list of statements. They were then asked, "Based on the reasoning given, which of the following do you think corresponds to the student's concluding statement?" This again took the form of a multiple-choice question and students were subsequently asked to explain their choice in a text box. Students were also asked, "Do you feel the student made any errors or omissions in the line of reasoning provided?" They were provided with a text box for their response and encouraged to reference specific lines in the fictitious student's argument by line number. Finally, in later iterations of the tasks, students were reminded of the initial answer they had given and asked if they wished to answer differently after examining the fictitious student's reasoning. For each base task, we developed at least two different fictitious reasoning chains: one that was correct and complete, and one that was incorrect. In designing the incorrect reasoning chains, we drew upon the physics education research literature and our own experience as educators to create a chain that would lead to a common incorrect response to the base task. As described below, students completing these assignments would be served only one of these fictitious chains.

For this investigation, four tasks have been administered using this basic format in calculus-based introductory physics sequences at different institutions, as described in Sec. II B. Two tasks, one concerning Newton's Laws and

the other concerning the work-energy relation, were administered in introductory mechanics courses. The other two tasks concern electric circuits and magnetic induction and were administered in introductory courses that covered those topics. All four base tasks were anticipated to be challenging for students, with fewer than half of the students expected to provide correct answers with correct reasoning on their own. This allowed us to gauge the ability of students to follow the reasoning that they did not spontaneously generate on their own. The tasks, and the fictitious student reasoning chains associated with each task, are described in detail below.

1. Follow task 1: Work and kinetic energy

The *work and kinetic energy* follow task, shown in Fig. 1, was drawn from the research literature in physics education [11,24,25]. The base task, shown in Fig. 1(a), involves two carts being pushed across a level surface of negligible friction. The carts have different masses ($m_A < m_B$) but are pushed by constant forces of the same magnitude F_o. The force is exerted on each cart only as the cart moves between the two marks on the table. Students were asked, "Is the kinetic energy of cart A *greater than, less than*, or *equal to* that of cart B after the carts have passed the second mark?"

Three fictitious student responses were created for this question. They are shown in Figs. 1(b)–1(d). The correct chain of reasoning is shown in Fig. 1(b). This reasoning chain supports the correct conclusion that the kinetic energy of cart A is equal to the kinetic energy of cart B after both carts have passed the second mark. The reasoning chain, shown in Fig. 1(c), referred to as the *v-wins chain*, is based on a statement made in a student dialogue from the tutorial "changes in energy and momentum" from *Tutorials in Introductory Physics* [11]. The v-wins chain supports the conclusion that the cart with a greater velocity, cart A, will have a greater kinetic energy after the two carts pass the second mark. The final fictitious student reasoning chain, shown in Fig. 1(d), was developed later and based on the most prevalent incorrect response observed in our student populations (as described in detail in Sec. IV A). Referred to as the *mass-only chain*, it supports the conclusion that the cart with more mass will have more kinetic energy (i.e., the kinetic energy of cart A will be less than the kinetic energy of cart B).

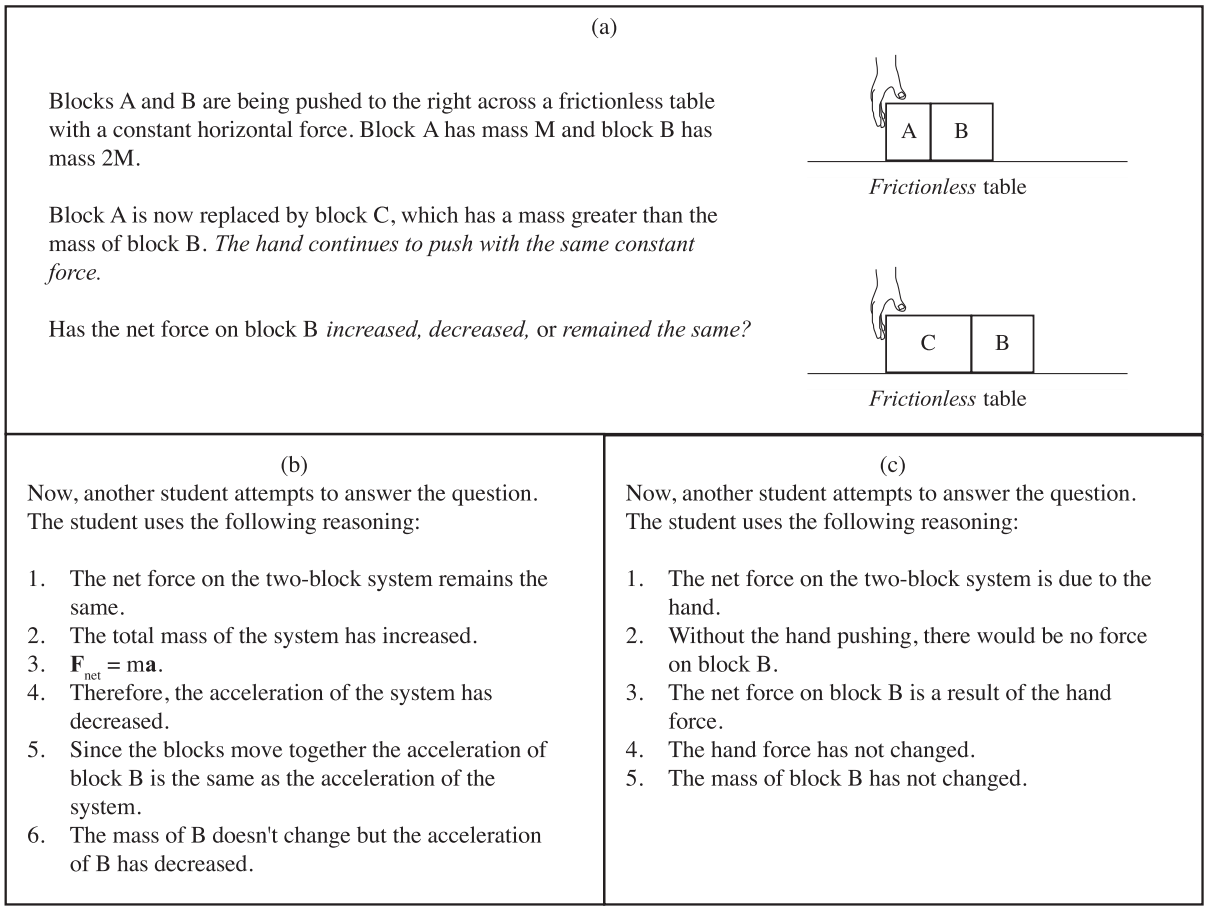


FIG. 2. Follow task 2, the *Newton's laws* follow task. The base task is shown in (a). The reasoning chain shown in (b) supports the correct conclusion that the net force on block B has decreased. The reasoning chain shown in (c) is incorrect and suggests that the net force on B has not changed.

2. Follow task 2: Newton's laws

The *Newton's laws* follow task was adapted from a task that has previously been used to assess student understanding of Newton's 2nd and 3rd laws [26]. The base task is shown in Fig. 2(a). It involves two blocks, A and B, being pushed across a surface of negligible friction by a hand exerting a constant force. Block A is replaced by a block of greater mass while the force exerted by the hand remains unchanged and students were asked what happens to the net force on block B. To respond correctly, students must realize that as the mass of the two-block system increases, the acceleration of the system must decrease. This in turn means that the net force on block B must decrease. This reasoning chain is the one used by the fictitious student response shown in Fig. 2(b). The fictitious student reasoning chain shown in Fig. 2(c) is based on a common error made on related exam questions; this chain stems from the incorrect assumption that the hand force is solely accountable for the net force on block B and thus leads to the conclusion that the net force on block B would not change.

3. Follow task 3: Electric circuits

The *electric circuits* follow task, shown in Fig. 2, involves qualitative reasoning about parallel circuit branches and resistance in simple electric circuits. The base task involves two circuits constructed from identical batteries, identical bulbs, and two boxes, X and Y, which have equal resistances. As shown in Fig. 2(a), students were asked to compare the brightness of bulbs 1 and 2. The correct response, that bulb 2 will be brighter than bulb 1, is supported by reasoning such as the chain shown in Fig. 3(b). The fictitious student reasoning chain, shown in Fig. 3(c), is one that we have commonly observed on similar exam questions; this chain suggests that the brightness of bulb 2 is equal to that of bulb 1.

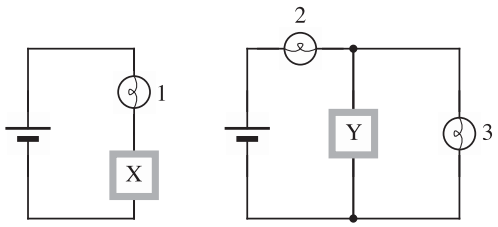
4. Follow task 4: Magnetic induction

The *magnetic induction* follow task requires students to apply Lenz's law to determine the direction of current in a copper loop placed in a region where the magnetic field strength is decreasing in time. The base task and the associated fictitious student responses are shown in Fig. 4. The correct reasoning invokes Lenz's law to argue that the induced field would be in the same direction as the external field in this case, and hence the current would flow around the ring counterclockwise as viewed from above. The fictitious student reasoning, shown in Fig. 4(b), represents this correct analysis. The student reasoning, shown in Fig. 4(c), reflects the common error that the induced field

(a)

In the circuits at right, all bulbs are identical, and the batteries are identical and ideal. Boxes X and Y contain unknown arrangements of bulbs. The resistances of boxes X and Y are equal, $R_x = R_y$.

Rank the brightness of bulbs 1 and 2.



Circuit I Circuit II

(b)

Now, another student attempts to answer the question. The student uses the following reasoning:

1. Adding an element in parallel decreases the resistance of a circuit.
2. Since Circuit II is similar to Circuit I with an additional bulb in parallel, it will have less resistance than Circuit I.
3. The battery in Circuit II will produce more current than the battery in Circuit I.
4. Bulbs 1 and 2 each receive all the current from the battery.
5. The brightness of a bulb is an indication of how much current flows through it.

(c)

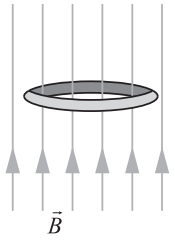
Now, another student attempts to answer the question. The student uses the following reasoning:

1. In each case, draw a loop that includes the battery, one bulb, and the box.
2. Each loop contains an identical bulb in series with an identical box.
3. The current doesn't split before reaching either bulb 1 or bulb 2.
4. Bulbs 1 and 2 each receive all the current from the battery.
5. The brightness of a bulb is an indication of how much current flows through it.

FIG. 3. Follow task 3, the *electric circuits* follow task. The base task is shown in (a). The reasoning chain shown in (b) supports the correct conclusion that bulb 2 will be brighter than bulb 1. The reasoning chain shown in (c) is incorrect and suggests that the two bulbs will be equally bright.

(a)

A copper wire loop is stationary in an external magnetic fields as represented by the magnetic field lines shown at right. The magnitude of the external magnetic field is **decreasing.**



Which choice best represents the direction of the current induced in the loop?
(a) There is no current in the loop.
(b) The current is flowing to the right across the front of the loop (i.e., counter -clockwise as viewed from above).
(c) The current is flowing to the left across the front of the loop (i.e., clockwise as viewed from above).

(b)

Now, another student attempts to answer the question. The student uses the following reasoning:

1. I choose an area vector upward, in the direction of the field lines.
2. Flux through the ring is positive and decreasing.
3. Therefore, the change in flux through the ring is negative.
4. By Lenz's law, the induced flux in the ring must be positive.
5. Therefore, the induced field must point in the same direction as the area vector.

(c)

Now, another student attempts to answer the question. The student uses the following reasoning:

1. I choose an area vector upward, in the direction of the field lines.
2. Flux through the ring is positive.
3. The induced flux in the ring must oppose the external flux.
4. Therefore, the induced flux in the ring must be negative.
5. Therefore, the induced field must point in the opposite direction to the area vector.

FIG. 4. Follow task 4, the *magnetic induction* follow task. The base task is shown in (a). The reasoning chain shown in (b) supports the correct conclusion that the induced current will flow counterclockwise in the loop. The reasoning chain shown in (c) is incorrect and suggests that the induced current will flow clockwise in the loop.

must oppose the external field, resulting in the incorrect conclusion that the induced current is clockwise as viewed from above.

B. Task administration and student population

Data for this study were collected in the calculus-based introductory physics courses [covering mechanics and/or electricity and magnetism (E&M)] at three universities in the United States, over a period of several semesters. Key similarities and differences between these institutions are noted in Table I. University A is a midsized, PhD-granting, public land-grant university in the Northeast. University B is a large, PhD-granting, public land-grant university in the Pacific Northwest. University C is a midsized, masters-granting, public university in the Pacific Northwest. All three universities serve student populations that are primarily white and the populations in the physics courses in

TABLE I. Key features and differences between the three universities where data were collected.

University	Highest degree granted	Size	Average math SAT score relative to national average	Use of tutorials	Administration of tasks	Assignment to task versions	Tasks administered
A	Ph.D.	Mid-sized	Slightly higher	Required recitation	Exam review	Random	Work-energy, Newton's laws, Eeectric circuits, magnetic induction
B	Ph.D.	Large	Substantially higher	Required recitation	Pretest for later tutorial	Pseudo-random	Work-energy, electric circuits, magnetic induction
C	Masters	Mid-sized	Slightly higher	Part of required lab	Exam review	Random	Work-energy, Newton's laws

which data were collected were predominantly male students intending to major in the sciences or engineering. Based on scores on the math SAT (a standardized test commonly used in college admissions), incoming first-year students at universities A and C demonstrate levels of mathematics preparation that are slightly higher than the national average [27], while at university B, the average Math SAT score for incoming students is substantially higher than the national average [28]. Not all tasks were administered at all three universities but rather were selected for administration based on course coverage and access (we did not have access to data collection in the course covering topics in E&M at university C). Each task was administered after all relevant instruction including the relevant tutorial from *Tutorials in Introductory Physics* or an adaptation thereof (at university C, the tutorials had been adapted to serve as portions of longer laboratory sessions). All tasks were administered out of class using an online survey tool. At universities A and C, each task was part of a longer online assignment that served as an exam review. At university B, each task was administered as part of an online pretest for a tutorial later in the quarter. In all cases, scoring on the assignment was entirely participation based. Students received a small amount of credit for completing the online assignment. Since each task included multiple versions, students were assigned to a particular version either randomly by the online survey tool Qualtrics [29] (at universities A and C) or pseudorandomly (at university B). In the pseudorandom condition, students were assigned to a survey version based on their recitation section. Although all students in each recitation section attended the same lecture section, care was taken that recitation sections associated with different lecture sections were reasonably uniformly distributed across the survey versions. In cases for which the same task was administered in different lecture sections at the same university in the same semester, data from multiple lecture sections have been combined (this happened for the work-energy task at universities B and C and for the Newton's laws task at university C). In all other cases, data have not been combined except where noted.

C. Data analysis

In each case, students first responded to the question as presented in the base task, and provided their own reasoning, before being presented with the fictitious student reasoning. Student answers were coded as correct or incorrect, and if incorrect, whether they aligned with the answer expected from one of the fictitious student reasoning chains. Reasoning was coded as to whether it was correct and complete.

In coding the reasoning of students giving a correct response to the base task, most responses were unambiguous. On the work-energy task, any mention of a comparison of the work done on the two carts was coded as correct and complete reasoning. On the Newton's laws task, students responses that made mention of the decreased acceleration of the system were coded as using correct and complete reasoning. On the electric circuits task, any mention of circuit 2 having a smaller resistance was coded as correct and complete reasoning. On the magnetic induction task, reasoning was coded as correct and complete if it mentioned opposing a change in field or flux. In a few cases, student reasoning was ambiguous or incomplete: for instance, on the magnetic induction task, many students gave their reasoning simply as "Lenz's law" without providing any further justification. While Lenz's law is certainly necessary for arriving at a correct response, this reasoning was judged insufficient and thus was not coded as aligning with either of the fictitious students. If a student had not provided any reasoning whatsoever (either by leaving the field blank or by typing "I don't know" or words or characters seemingly unrelated to the task at hand), they were deemed to have not taken the task seriously and their responses were removed from subsequent analysis. (Typically 5%–10% of students who had provided an answer to the base task were removed for this reason.) Students who typed a relevant equation but nothing more, such as "$\mathrm{KE} = \frac{1}{2}mv^2$" for the work-energy task, or the words Lenz's law for the induction task, were still considered to have provided reasoning.

In coding the incorrect responses, on three of the tasks, student reasoning was fairly limited, and thus reasoning given by the students with incorrect answers was not coded in detail; rather it was noted whether or not their answer aligned with the answer given by the incorrect student. For the work and kinetic energy task, however, students used a wide range of ideas to arrive at each of the available incorrect answers, and thus the reasoning of students giving incorrect responses was also coded as to whether it aligned with the reasoning provided by the incorrect fictitious student, or whether it was based on ideas not articulated by any of the fictitious students.

On the work and kinetic energy task, students were coded as aligning with the "mass only" chain if they used the correct equation for kinetic energy but only mentioned a comparison of the masses of the two carts: "kinetic energy is $\frac{1}{2}m(v)^2$ and cart A has a lower mass than cart B." They were coded as aligning with the "v-wins chain" if their answer gave an indication that velocity matters more in the equation $\mathrm{KE} = \frac{1}{2}mv^2$ than mass does, for example, "KE is equal to half the mass times velocity squared and Force = mass $\times$ acceleration. Because [of] the force applied cart A will have a higher acceleration meaning it will have a higher velocity. And since KE squares velocity, cart A will in most cases will have a higher KE." If a student made no mention of the

velocity being squared (outside of the equation) but mentions both mass and velocity in their explanation, they were coded as not aligning with any of the fictitious students: "The carts are not moving at the same velocities. While both of those carts were given the same force, cart A will move faster as it has less mass. The way to evaluate kinetic energy is ½ mv^2. If the mass of cart A is less, but the velocity is greater, and the mass of cart B is greater, but the velocity is less, those 2 balance each other out."

Each student was also asked to select a conclusion for the fictitious student reasoning with which they were presented; this selection was also coded as correct or incorrect based on whether or not it aligned with the appropriate conclusion for that reasoning chain. Student explanations for why they had selected this conclusion to the fictitious chain and their critiques of the chain were, in general, not coded in detail. However, to ensure that students were engaging with the task as intended, one sample of students responding to the correct-reasoning version of the electric circuits task at university A was examined more carefully. For this one sample, we coded student responses for whether or not they demonstrated engagement with the reasoning chain in either the textbox in which they were asked to explain their reasoning for their choice of the fictitious student's closing statement or in the textbox in which they were asked to critique the fictitious student's statement. To be coded as having demonstrated engagement with the reasoning chain, students had to reference at least one specific statement in the chain or had to provide a reasonable summary of the full chain. Examples of responses that were coded as demonstrating engagement with the chain include "the student's statement of bulbs 1 and 2 receiving the current from the battery," "the student thinks that the second circuit produces more current in the battery, and all current flows through bulb 1 in circuit 1 and bulb 2 in circuit 2, but since the current in circuit 2 is greater, more current will be flowing through bulb 2. And since power (or brightness) is proportional to how much current flows through a bulb, then bulb 2, the bulb with more current will be brighter," and, as a critique, "we know that the batteries are identical to one another, so the current produced by them should be equal, unlike point 3 made by the student." Examples of responses that were coded as not providing evidence for engagement with the reasoning chain include "they have the same current flowing through them," "more current flows through bulb 2," and "less resistance means more current." In some of these cases, however, it is possible that the respondent was well engaged with the fictional reasoning chain, thus we believe that our coding provides a conservative estimate of student engagement with the chain.

Fisher's exact test was used in all cases where statistical analyses were needed. Fisher's exact test determines whether differences between the percentage of students responding correctly in two populations (in this case, the population of students who gave a correct response with correct reasoning for the base task vs those who did not) are unlikely to be due to chance alone and is an appropriate test when the outcome variable (in this case, whether a student's response is correct or incorrect) is categorical. Fisher's exact test is used as an alternative to the chi-square test when N values are lower because it provides an exact measure of the probability rather than an approximation. The assumptions of Fisher's exact test are that (i) observations are independent, and (ii) observations are mutually exclusive (i.e., each observation can fall into only one cell of a contingency table). Both of these hold for our data.

IV. RESULTS

A. Performance on base tasks

As intended, the questions presented in each base task proved challenging for students. Student performance on each base task is shown in Figs. 5–8. The proportion of students giving correct answers (both with and without correct reasoning), as well as answers that align with the fictitious student responses for each task are shown. Rates of correct reasoning were typically 40% or lower. For tasks 1–3, the common response aligned with one of the incorrect fictitious student responses. On the magnetic induction task (task 4), however, the most common response was the correct response. Many of these correct answers on the magnetic induction task were not supported with correct and complete reasoning. (As noted earlier, many students gave their reasoning simply as Lenz's law or "right-hand rule" without explaining how they had applied these rules. These responses, which were used to support both correct and incorrect responses, were not coded as representing correct reasoning.)

B. Ability to follow reasoning chains

Students were generally successful at following each of the fictitious reasoning chains. In each case, more than half of the students followed the reasoning chain correctly; in most cases, it was more than 2/3 of students who followed the chain correctly, regardless of whether the chain represented correct or incorrect reasoning, as shown in Table II. In the one section for which student explanations for their choice were examined in detail (a section of students responding to the correct reasoning chain for the electric circuits task at university A), the majority of students, 64%, provided explanations or critiques for the fictitious student's argument that suggested engagement with the physics of the argument, supporting the idea that students

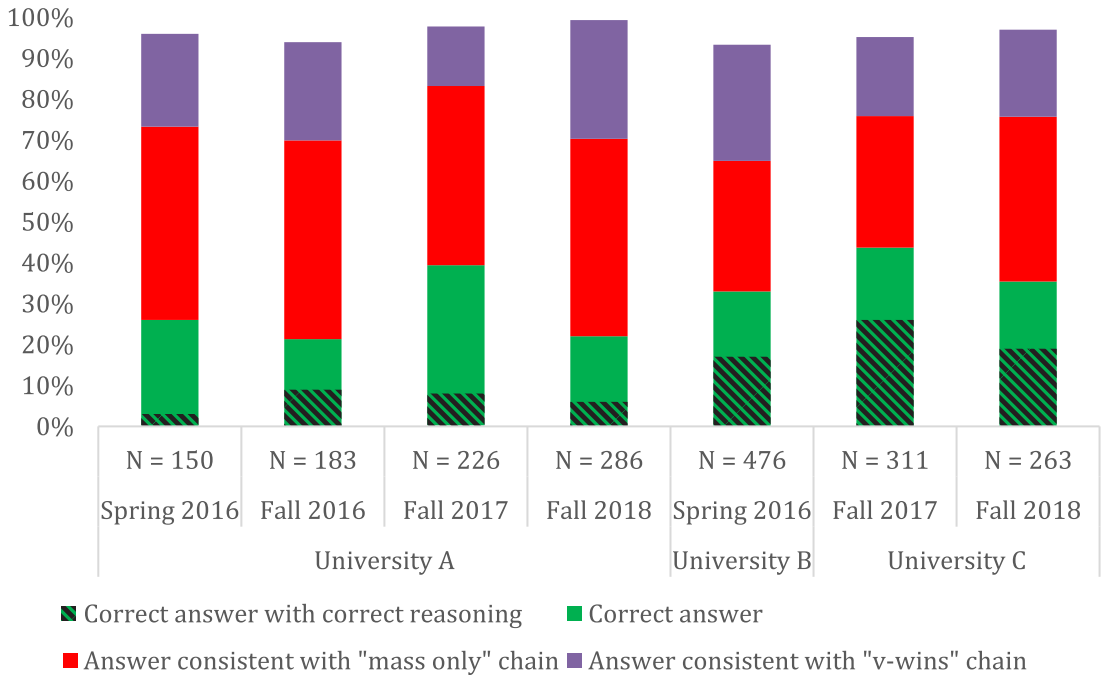


FIG. 5. Student responses on the base work-energy task. Shaded areas represent correct answers supported with correct reasoning. All other regions represent answers only, irrespective of the reasoning given. Totals do not sum to 100% because answer choices not aligned with one of the fictitious students are not included.

are not just selecting the correct conclusion for the chain based on its surface features. Success at inferring appropriate conclusions from a qualitative inferential reasoning chain had previously been demonstrated at a single institution for fairly straightforward reasoning chains [14], but here we demonstrate that the ability to follow chains of reasoning holds for more conceptually challenging chains of reasoning across a variety of topics and institutions.

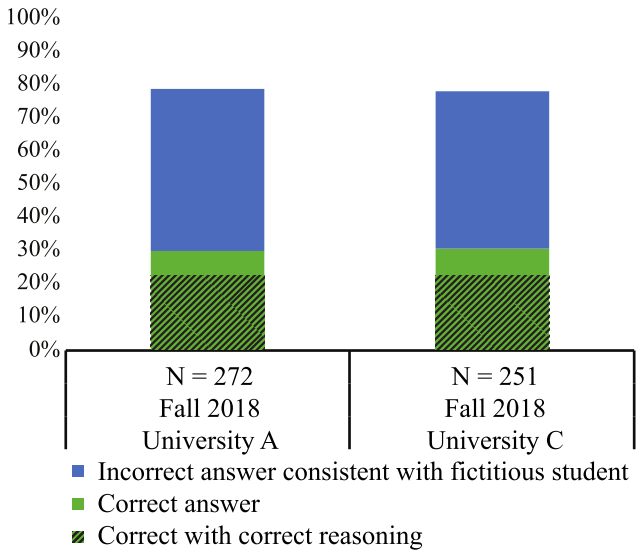


FIG. 6. Student responses on the Newton's laws base task. Shaded areas represent correct answers supported with correct reasoning. All other regions represent answers only, irrespective of the reasoning given. Totals do not sum to 100% because answer choices not aligned with one of the fictitious students are not included.

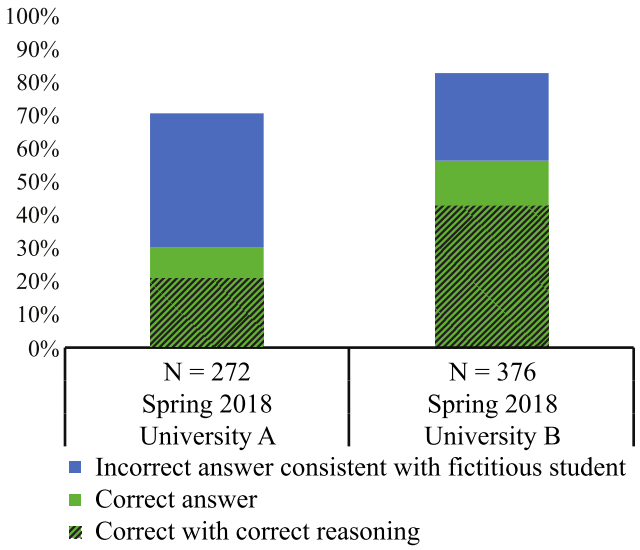


FIG. 7. Student responses on the electric circuits base task. Shaded areas represent correct answers supported with correct reasoning. All other regions represent answers only, irrespective of the reasoning given. Totals do not sum to 100% because answer choices not aligned with one of the fictitious students are not included.

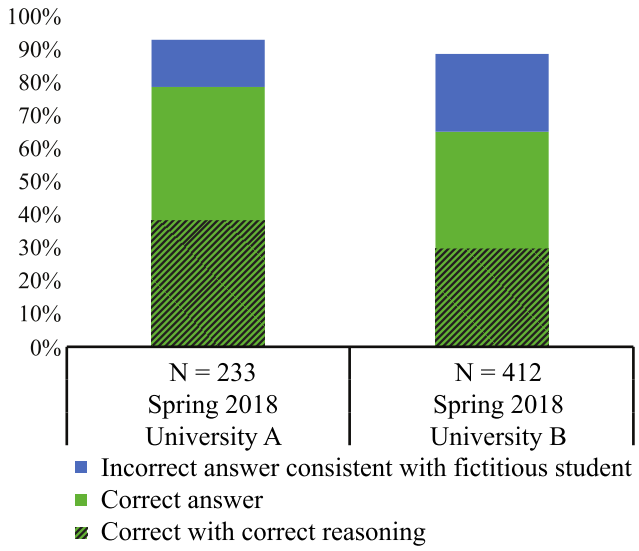


FIG. 8. Student responses on the magnetic induction base task. Shaded areas represent correct answers supported with correct reasoning. All other regions represent answers only, irrespective of the reasoning given. Totals do not sum to 100% because answer choices not aligned with one of the fictitious students are not included.

C. Interactions between performance on base tasks and ability to follow reasoning chains

Student ability to follow the reasoning chains appears less uniform when individual performance on the base tasks is taken into account. Table III indicates the fraction of students who followed the chain correctly among those who had given a correct answer and provided correct reasoning on the base task (referred to as "CwCR" students) as well as the fraction of students who followed the chain correctly among all other students. In all cases, Fisher's exact test was used to determine whether CwCR students were more successful at following the chain than students who had not used correct reasoning or had not given a correct response. On the correct reasoning chains, CwCR students were uniformly more successful at inferring the appropriate reasoning from the chains than other students (p values calculated using Fisher's exact test are shown in Table III).

On the incorrect reasoning chains, students who had not responded correctly or used correct reasoning were typically as successful as the CwCR students, with one exception. That one exception was the magnetic induction task, on which CwCR students were substantially more likely than other students to select the appropriate conclusion for the incorrect reasoning chain. This may be because for many students, answering a question using Lenz's law correctly requires the application of algorithmic reasoning. We speculate that students who were able to generate this algorithm correctly on their own may also have been more successful at stepping through an alternate algorithm such as the one presented by the incorrect fictitious student. However, this result may also be linked to the high "false positive" results on the magnetic induction base task (many students gave a correct response but did not support it with correct reasoning). It is possible that these students have begun to develop the mindware associated with Lenz's law, but if this mindware is not sufficiently well developed that they can articulate it clearly, they may also not be ready to recognize the distinction between a correct application of Lenz's law and the incorrect reasoning shown in Fig. 4(c).

It might be expected that students who themselves had used reasoning similar to the incorrect student would be

TABLE II. Average percentage of students who selected the appropriate conclusion when presented with a particular chain, at each institution. Note that students were randomly served only one of the possible reasoning chains, thus each entry in the table represents a distinct group of students. In some cases, data from multiple semesters and lecture sections have been combined. Not all tasks were administered at each institution. Students who did not provide reasoning when responding to the question posed in the base task were omitted from this analysis.

		% of students selecting the appropriate conclusion to chain		
		Institution		
Chain		A	B	C
Work-energy	Correct reasoning	82% ($N = 373$)	80% ($N = 130$)	87% ($N = 286$)
	v-wins (incorrect)	81% ($N = 131$)	…	94% ($N = 120$)
	Mass only (incorrect)	86% ($N = 103$)	…	92% ($N = 153$)
Newton's laws	Correct reasoning	65% ($N = 97$)		73% ($N = 114$)
	Incorrect reasoning	91% ($N = 137$)		95% ($N = 130$)
Electric circuits	Correct reasoning	56% ($N = 101$)	88% ($N = 164$)	
	Incorrect reasoning	91% ($N = 108$)	91% ($N = 171$)	
Magnetic induction	Correct reasoning	73% ($N = 107$)	79% ($N = 155$)	
	Incorrect reasoning	68% ($N = 115$)	73% ($N = 190$)	

TABLE III. Student ability to follow each reasoning chain, broken out by performance on the base task. "CR" and "IR" refer to correct and incorrect reasoning chains, respectively. In some cases, data from multiple semesters and lecture sections have been combined. *P* values were calculated using Fisher's exact test to test whether students who had given a correct answer and provided correct reasoning (CwCR) were more likely to follow the reasoning chain correctly than other students. The "other" group includes both students who had given an incorrect answer on the base task as well as those who had given a correct answer on the base task but had provided an incorrect justification. Statistically significant differences are indicated in boldface.

		Institution								
		A			B			C		
Chain		Among CwCR	Among others	p	Among CwCR	Among others	p	Among CwCR	Among others	p
Work-energy	CR	**100% (*N* = 23)**	**80% (*N* = 350)**	**0.011**	**100% (*N* = 25)**	**75% (*N* = 105)**	**0.004**	**97% (*N* = 70)**	**84% (*N* = 216)**	**0.003**
	v-wins (IR)	80% (*N* = 10)	81% (*N* = 121)	1				100% (*N* = 24)	92% (*N* = 93)	0.342
	M-only (IR)	80% (*N* = 10)	87% (*N* = 93)	0.623				97% (*N* = 39)	91% (*N* = 111)	0.29
Newton's laws	CR	**100% (*N* = 21)**	**55% (*N* = 76)**	**<0.001**				**90% (*N* = 29)**	**67% (*N* = 85)**	**0.028**
	IR	88% (*N* = 40)	93% (*N* = 97)	0.332				93% (*N* = 20)	96% (*N* = 103)	0.604
Electric circuits	CR	**86% (*N* = 22)**	**48% (*N* = 79)**	**0.001**	**96% (*N* = 77)**	**82% (*N* = 87)**	**0.006**			
	IR	93% (*N* = 27)	90% (*N* = 81)	1	93% (*N* = 93)	90% (*N* = 79)	0.589			
Magnetic induction	CR	**93% (*N* = 42)**	**60% (*N* = 65)**	**<0.001**	**95% (*N* = 64)**	**71% (*N* = 133)**	**<0.001**			
	IR	**85% (*N* = 48)**	**55% (*N* = 67)**	**0.001**	**92% (*N* = 61)**	**64% (*N* = 129)**	**< 0.001**			

more successful at following the incorrect chains rather than those students who had responded correctly with correct reasoning. Although we have not shown our data broken out in this way, we did conduct such an analysis and observed that our data are indeed suggestive of such an alignment bias for three of the four tasks (i.e., students were more likely to infer that each fictitious student would arrive at the same conclusion they themselves had given, with the exception of the magnetic induction task). However, we do not have sufficient statistical power to demonstrate this effect with any one population. We thus cannot conclude definitively that students who used the same reasoning as the incorrect fictitious student were any more or less successful at following that reason than were students who had reasoned correctly for themselves.

V. DISCUSSION AND CONCLUSIONS

Mounting evidence suggests that students may not always reason on the basis of first principles and chain ideas together until reaching a conclusion [1,17,30]. This phenomenon may be based in part on the nature of human reasoning itself [1] and raises questions as to the effectiveness of instruction that does emphasize logical reasoning chains, such as many research-based instructional materials like *Tutorials in Introductory Physics* [11]. We have found that, in general, students are capable of correctly interpreting and inferring the appropriate conclusion when presented with a qualitative inferential reasoning chain. In other words, when presented with the solution to a qualitative problem that uses step-by-step reasoning to explain its solution, our findings suggest that most students will follow the reasoning chain correctly. While this had previously been demonstrated at one institution for fairly simple tasks [14,31], here we have shown that this finding holds at multiple institutions, across content domains, and for tasks on which students are considerably less likely to generate the correct chain on their own. This finding is particularly important given the prevalence of qualitative inferential reasoning chains in instructional materials such as the student dialogues in tutorial worksheets or the solutions to example problems, qualitative exam questions, and homework problems provided to students in many introductory physics courses.

We have also demonstrated, however, that students who do not generate a correct chain on their own are significantly less likely to follow the correct chain to its appropriate conclusion than those students who did

independently generate correct reasoning. This finding was drawn from data from multiple universities and across content domains. While in most cases, 2/3 or more of the students who responded incorrectly on the base task nonetheless were able to follow the correct reasoning chain, this was not uniformly true. In some cases, more than 40% (and in one case, more than 50%) of students who had responded incorrectly to the task inferred an inappropriate conclusion when presented with a correct reasoning chain. Thus, the students who stand to benefit most from reading and reflecting upon the correct solution to a physics question are less likely to actually recognize what answer it is suggesting. Whether this is due to a shortcoming of the students' mindware, their cognitive reflection skills, a combination of these factors, or something else entirely remains to be investigated. In any case, instructors should be aware that students are not always drawing appropriate conclusions from the solutions they present. This finding may help to explain why certain incorrect student reasoning patterns persist even after completing targeted research-based instructional materials [32]. Indeed, such students might benefit from additional instructional time devoted to strengthening their reasoning skills. For these students, an increased instructional emphasis on the construction, analysis, and evaluation of reasoning chains might be especially beneficial.

Students also tended to be quite successful at following the common incorrect reasoning chains. In many cases, more than 85% of students followed the chain correctly, regardless of whether they themselves had used correct or incorrect reasoning when responding to the base task. This suggests that most students will follow incorrect reasoning chains laid out in, for example, student dialogues in instructional materials. This finding may be related to the strong intuitive appeal of the particular responses shown, in accordance with dual-process theories of reasoning. This finding also has implications for the development of pedagogical content knowledge for future teachers [33,34], suggesting that they can be successful at following common lines of reasoning articulated by many students and anticipating the conclusion that their students will infer from this reasoning, even if they are still developing their own content understanding.

For the electric circuits task, on which we coded student reasoning about the conclusion to the fictitious chain in detail, we noted that student responses to why they had chosen a particular concluding statement seemed to emphasize different elements of the chain depending on the concluding statement they selected for the chain. Students who indicated that the fictitious student would choose "$2 < 1$" tended to focus on the student's mention of the parallel connection. Students who indicated that the fictitious student would choose $2 = 1$ frequently mentioned the statement that "bulbs 1 and 2 each receive all the current from the battery." This selective focus on statements that students can align with their chosen conclusion is consistent with the dual-process theories framework, which suggests that students might tend to preferentially search for evidence that supports their first-available conclusion rather than being on the lookout for information that invalidates their conclusion. Prior studies have shown similar effects [9,35].

VI. LIMITATIONS

The follow tasks were designed to mimic typical written explanations provided to students on tutorial worksheets and exam or homework solutions. Incorrect reasoning chains were included to represent the "student dialogues" or "student statements" common to many research-based instructional materials. The written nature of these tasks, however, does present some limitations. It is possible that students were misinterpreting the fictitious student arguments as they were written and would have followed an oral explanation, as in a conversation with another student, more accurately.

It is also possible that students merely skimmed the fictitious student arguments and identified a plausible concluding statement without engaging deeply with the physics of the argument being made; however, the fact that, for the one task we examined in detail, nearly 2/3 of students were able to provide a substantive commentary on or critique of the fictitious student's reasoning suggests that students were engaging with the task at a level deeper than just skimming. We cannot definitely rule out, however, that students would be perfectly capable of following the arguments made by the fictitious students in another context and simply chose not to when the task is part of a participation-credit online homework assignment.

Although our data collection spanned multiple universities, our population for this study was primarily students who are white, male, and with higher-than-average mathematics preparation. Further study is needed to determine the extent to which these results hold for other student populations.

Finally, we recognize that not all readers may agree with the operational definition of following a reasoning chain (selecting the appropriate conclusion when presented with a logical chain of reasoning elements) that we have used in this work. We feel, however, that whatever name it is given, the skill we are measuring—that of inferring an appropriate conclusion based on an argument made by someone else—is a useful skill for students and one that is frequently expected of the students in our courses.

VII. IMPLICATIONS FOR INSTRUCTION AND FUTURE RESEARCH

The work presented here is an initial exploration that opens up many intriguing avenues for future investigation. In the present work, for instance, we did not attempt to

distinguish between students using anything other than the reasoning they had generated for the particular task on which we were assessing them. However, it is possible that some other characteristics could be used to identify which students are more likely to infer the appropriate conclusion from a reasoning chain. According to dual-process models of human cognition, in order to respond correctly to the base task, students must have both an appropriate level of mindware and a sufficient tendency toward cognitive reflection to allow for a sustained override of any incorrect intuitive responses [18]. Indeed, several studies have shown that students with a greater tendency toward cognitive reflection, as measured by the CRT, are also less likely to give an intuitively appealing incorrect answer to certain qualitative questions [22] or are more likely to benefit from certain research-based instructional interventions [23]. Thus, we would expect that students with a higher CRT score would also be more likely to respond correctly to the base tasks presented in this work. It also seems likely, however, that the tendency toward cognitive reflection would also lead students to reflect more carefully on a reasoning chain presented to them. Thus even if a student does not have sufficient mindware to generate a chain on their own, we predict that those students with a greater tendency toward cognitive reflection might be more likely to detect a conflict between their own response and the reasoning presented to them in the fictitious student statements and would be better able to sustain an override of their initial response in order to follow the presented chain correctly. In other words, even among students who responded incorrectly to the base task themselves, we predict that those with a greater tendency toward cognitive reflection would be more likely to infer the correct conclusion to the fictitious student reasoning chains. Studies that measure both students' CRT scores and their ability to follow a reasoning chain, and examine associations between these variables, are needed to verify this prediction.

We suspect that the tendency toward cognitive reflection may also influence how much instructional benefit a student derives from completing the follow tasks. We are particularly interested in those students who respond incorrectly when initially presented with the base task, but then go on to follow the correct reasoning chain appropriately. When reminded of their original answer and asked if they would like to change their own response, only about 25% of these students choose to revise their answer. The other students continued to endorse their original answer by selecting, "No, I still agree with my original answer." We suspect that for students with a greater tendency toward cognitive reflection, the correct reasoning presented in the follow tasks may provide a sufficient "nudge" that they may be able to respond correctly to the base task after reading and analyzing the fictitious student's reasoning chain. We speculate that for students with a lower tendency toward cognitive reflection, the act of reading a correct solution, even if they follow this solution and recognize the conclusion it implies, may not provide sufficient impetus for them to critically examine and override their own response. A more detailed analysis of how student responses to the base task change after interacting with the fictitious students' reasoning chains is the subject of ongoing work.

The results of this study may also suggest a viable path forward for efforts designed to strengthen student reasoning in physics. At a minimum, we hope that interacting with the fictitious student chains, particularly those representing correct and complete reasoning, may improve students' topic-specific content understanding. It is also possible, however, that extensive exposure to such reasoning activities, which require students to engage in sustained mental simulation in which they set aside their own ideas and explore other arguments, may help students to build the skills needed for conflict detection and sustained override. We speculate that practicing with tasks that require students to infer the appropriate conclusion to a given reasoning chain may encourage cognitive reflection and might, ultimately, result in the improvement of student reasoning skills. Indeed, we are currently exploring the viability and impact of follow tasks (and modified follow tasks) as instructional interventions as part of a larger project focused on research-based curriculum development efforts aligned with dual-process theories of reasoning.

ACKNOWLEDGMENTS

The authors are extremely grateful to Peter Shaffer, Saima Farooq, Kevin Covey, and other instructors of courses in which data were collected. The authors would also like to acknowledge substantive discussions with Mila Kryjevskaia, Paula R. L. Heron, and Andrew Boudreaux about this work. This material is based upon work supported by the National Science Foundation under Grants No. DUE-1821390, No. DUE-1821123, No. DUE-1821400, No. DUE-1821511, No. DUE-1821561, No. DUE-1431940, No. DUE-1431541, No. DUE-1431857, No. DUE-1432052, No. DUE-1432765, and No. DRL-0962805.

[1] M. Kryjevskaia, M. R. Stetzer, and N. Grosz, Answer first: Applying the heuristic-analytic theory of reasoning to examine student intuitive thinking in the context of physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020109 (2014).

[2] A. F. Heckler and A. M. Bogdan, Reasoning with alternative explanations in physics: The cognitive accessibility rule, Phys. Rev. Phys. Educ. Res. **14**, 010120 (2018).

[3] B. A. Lindsey, M. L. Nagel, and B. N. Savani, Leveraging understanding of energy from physics to overcome unproductive intuitions in chemistry, Phys. Rev. Phys. Educ. Res. **15**, 010120 (2019).

[4] L. C. McDermott, Guest Comment: How we teach and how students learn—a mismatch?, Am. J. Phys. **61**, 295 (1993).

[5] P. R. L. Heron, Empirical investigations of learning, and teaching, part II: Developing research-based instructional materials, in *Proceedings of the International School of Physics "Enrico Fermi" Course CLVI: Research on Physics Education*, edited by E. F. Redish and M. Vincentini (IOS Press, Varenna, Italy, 2003), pp. 351–365.

[6] E. F. Redish, Oersted Lecture 2013: How should we think about how our students think?, Am. J. Phys. **82**, 537 (2014).

[7] D. Kahneman, *Thinking, Fast and Slow* (Farrar, Strauss, & Giroux, New York, 2011).

[8] J. S. B. T. Evans and K. E. Stanovich, Dual-process theories of higher cognition, Perspect. Psychol. Sci. **8**, 223 (2013).

[9] J. C. Speirs, M. R. Stetzer, B. A. Lindsey, and M. Kryjevskaia, Exploring and supporting student reasoning in physics by leveraging dual-process theories of reasoning and decision making, Phys. Rev. Phys. Educ. Res. **17**, 020137 (2021).

[10] C. R. Gette, M. Kryjevskaia, M. R. Stetzer, and P. R. L. Heron, Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy, Phys. Rev. Phys. Educ. Res. **14**, 010113 (2018).

[11] L. C. McDermott, P. S. Shaffer, and the Physics Education Group at the University of Washington, *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).

[12] L. C. McDermott and the Physics Education Group at the University of Washington, *Physics by Inquiry* (John Wiley & Sons, New York, 1996).

[13] F. M. Goldberg, V. K. Otero, and S. Robinson, *Physics and Everyday Thinking*, 2nd ed. (It's About Time, Armonk, NY, 2007).

[14] W. N. Ferm Jr., J. C. Speirs, M. R. Stetzer, and B. A. Lindsey, Investigating student ability to follow and interact with reasoning chains, in *proceedings of PER Conf. 2016, Sacramento, CA*, 10.1119/perc.2016.pr.025.

[15] J. S. B. T. Evans, The heuristic-analytic theory of reasoning: Extension and evaluation., Psychon. Bull. Rev. **13**, 378 (2006).

[16] V. Talanquer, Chemistry education: Ten heuristics to tame, J. Chem. Educ. **91**, 1091 (2014).

[17] A. F. Heckler, The ubiquitous patterns of incorrect answers to science questions: The role of automatic, bottom-up processes, in *Psychology of Learning and Motivation—Advances in Research and Theory* (Elsevier Inc., New York, 2011), pp. 227–267.

[18] K. E. Stanovich, Miserliness in human cognition: The interaction of detection, override and mindware, Think. Reas. **24**, 423 (2018).

[19] K. E. Stanovich, *What Intelligence Tests Miss: The Psychology of Rational Thought* (Yale University Press, New Haven, CT, 2009).

[20] S. Frederick, Cognitive reflection and decision making, J. Econ. Perspect. **19**, 25 (2005).

[21] A. K. Wood, R. K. Galloway, and J. Hardy, Can dual processing theory explain physics students' performance on the Force Concept Inventory?, Phys. Rev. Phys. Educ. Res. **12**, 023101 (2016).

[22] C. R. Gette and M. Kryjevskaia, Establishing a relationship between student cognitive reflection skills and performance on physics questions that elicit strong intuitive responses, Phys. Rev. Phys. Educ. Res. **15**, 010118 (2019).

[23] M. Kryjevskaia, M. R. Stetzer, B. A. Lindsey, A. McInerny, P. R. L. Heron, and A. Boudreaux, Designing research-based instructional materials that leverage dual-process theories of reasoning: Insights from testing one specific, theory-driven intervention, Phys. Rev. Phys. Educ. Res. **16**, 020140 (2020).

[24] R. Lawson and L. C. McDermott, Student understanding of the work-energy and impulse-momentum theorems, Am. J. Phys. **55**, 811 (1987).

[25] T. O'Brien-Pride, S. Vokos, and L. C. McDermott, The challenge of matching learning assessments to teaching goals: An example from the work-energy and impulse-momentum theorems, Am. J. Phys. **66**, 147 (1998).

[26] 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).

[27] collegedata.com (2018).

[28] S. Kanim and X. C. Cid, Demographics of physics education research, Phys. Rev. Phys. Educ. Res. **16**, 020106 (2020).

[29] Qualtrics, http://www.qualtrics.com (2021).

[30] M. Kryjevskaia, P. R. L. Heron, and A. F. Heckler, Intuitive or rational? Students and experts need to be both, Phys. Today **74**, No. 8, 28 (2021).

[31] W. N. Ferm Jr., Examining student ability to follow and interact with qualitative inferential reasoning chains, Masters thesis, University of Maine, 2017.

[32] M. Kryjevskaia, M. R. Stetzer, and P. R. L. Heron, Student understanding of wave behavior at a boundary: The limiting case of reflection at fixed and free ends, Am. J. Phys. **79**, 508 (2011).

[33] L. S. Shulman, Those who understand: Knowledge growth in teaching, Educ. Res. **15**, 4 (1986).

[34] J. R. Thompson, W. M. Christensen, and M. C. Wittmann, Preparing future teachers to anticipate student difficulties in physics in a graduate-level course in physics, pedagogy, and education research, Phys. Rev. ST Phys. Educ. Res. **7**, 010108 (2011).

[35] M. L. Nagel and B. A. Lindsey, Implementation of reasoning chain construction tasks to support student explanations in general chemistry, J. Chem. Educ. **99**, 839 (2022).