Exploring and supporting student reasoning in physics by leveraging dual-process theories of reasoning and decision making

J. Caleb Speirs

Department of Physics and Astronomy and Maine Center for Research in STEM Education, University of Maine, Orono, ME 04469, USA

School of Mathematical and Physical Sciences, University of New England, Biddeford, ME 04005, USA

MacKenzie R. Stetzer

Department of Physics and Astronomy and Maine Center for Research in STEM Education, University of Maine, Orono, ME 04469, USA

Beth A. Lindsey

Department of Physics, Penn State Greater Allegheny, McKeesport, PA 15132, USA

Mila Kryjevskaia

Department of Physics, North Dakota State University, North Dakota, Fargo, ND 58108, USA

(Received 22 October 2019; revised 15 July 2021; accepted 23 September 2021; published 24 November 2021)

Because of the focus of introductory physics courses on improving students' problem-solving and reasoning skills, researchers in physics education have been developing and refining theoretical frameworks for how students reason through physics problems. Recently, researchers have begun to apply dual-process theories of reasoning (DPToR), from cognitive science and psychology, to support mechanistic predictions of student reasoning in physics. In this article, we employ a novel methodology involving reasoning chain construction tasks in order to test DPToR-based predictions for two physics questions in which salient distracting features have been found to cue incorrect first-available mental models. In a reasoning chain construction task, students respond to a physics question by drawing from a list of reasoning elements (all of which are true) in order to assemble a chain of reasoning that leads to a conclusion. Two sets of experiments were conducted to test the hypothesis that students would be unlikely to abandon an incorrect first-available model unless they were provided with information that called into question the satisfactoriness of that model. We found that providing increased access to information relevant to the correct line of reasoning did not produce large differences in student answering patterns. However, providing increased access to information refuting the first-available model did produce large differences in student answering patterns, but only among those students who demonstrated that they possessed the relevant mindware (i.e., conceptual understanding). Our findings are consistent with DPToR and further illustrate the applicability of such reasoning frameworks in the context of physics.

DOI: 10.1103/PhysRevPhysEducRes.17.020137

I. INTRODUCTION

Many students take introductory physics courses in service of other majors in a variety of different science, technology, engineering, and mathematics fields. It is often expected that these students will take the knowledge gained and, perhaps more importantly, the reasoning skills acquired in the course and employ them in their respective fields of study. Research-based instructional materials and approaches have been demonstrated to increase student conceptual understanding of core physics concepts [1,2], but little of this work has expressly explored the process of reasoning itself. Additionally, even after instruction using research-based approaches, it remains difficult to increase student performance on certain qualitative physics questions [3,4]. More detailed research into these questions has led physics education researchers to believe that processes generic to all human reasoning—that is, not necessarily associated with physics content—may be impacting the way students answer these questions [3–5]. As a result, many researchers have increasingly begun to investigate the cognitive mechanisms that influence human reasoning and how they affect student reasoning on qualitative physics questions [6–9].

Dual-process theories of reasoning (DPToR) have played a key role in a renewed effort to understand the mechanisms behind student reasoning. These theories arise from findings in cognitive science, social psychology, and the psychology of reasoning. Popularized by Kahneman [10], DPToR model human reasoning via two types of processing: (1) an unconscious, fast, and associative process; and (2) a conscious, effortful, and typically slower process. These theories tend to be mechanistic in nature; as such, they provide a framework that can easily be prescriptive and provide a basis for progress in developing successful instructional interventions.

While dual-process theories are useful for understanding domain-general cognitive mechanisms and their impact on student use of conceptual knowledge on a given physics problem, new research methodologies that can disentangle student reasoning skills from conceptual understanding are also needed. Our collaboration has sought to develop and refine such methodologies, and this paper presents one of these novel methodologies, which centers around the *reasoning chain construction task*. This methodology has been useful in studying explicit process 2 reasoning, especially the formation and structure of student's qualitative inferential reasoning chains.

Several research questions related to dual-process theories emerged from our ongoing work on reasoning chain construction tasks, which extends beyond the investigations described in this manuscript. Can reasoning chain construction tasks be used in order to explore the extent to which dual-process theories of reasoning successfully predict patterns in student reasoning and answer choices on certain physics questions? In particular, can reasoning chain construction tasks be used to examine aspects of these dual-process frameworks that have been previously untested in the context of physics?

Accordingly, in this paper, we draw upon dual-process theories to make predictions for student behavior on chaining tasks, including chaining tasks that contain modest interventions based on these theories. The findings from our work provide additional support for the reasoning mechanisms put forward by many dual-process theories and have implications for the development of instructional materials.

II. BACKGROUND AND MOTIVATION

When a student answers a qualitative physics question incorrectly, it is often assumed that the student did not possess a robust understanding of the physics involved. It is also commonly presumed that the student reasoned from an incorrect or incomplete conception of the relevant physics. There are differing perspectives as to the structure of these conceptions. One perspective is that physics (mis)conceptions, once learned, are stable and robust and that the same (mis)conception would be applied in every instance in which they are needed [11,12], much like a car, once manufactured, is used whenever one perceives that a car is needed. Another perspective [13–15] holds that physics conceptions are built from fragmentary knowledge and resources assembled at the time the task is being performed, much like a toy car assembled from toy construction bricks; as such, each conception is inherently unstable and can emerge as a slightly different structure based on the in-the-moment perception of the demands of a task. The former perspective is generally referred to as the “misconceptions” framework, while the latter is referred to as the “resources” framework. A third, alternate way of modeling student reasoning is to investigate student “difficulties”; in this perspective, the emphasis is not on the cognitive structure of the knowledge or its stability, but rather on the identifying characteristics of that knowledge and the frequency of its occurrence among a population of students [16–18], or, to continue the analogy, the percentage of students who use a specific type of car to solve a given problem.

One challenge within the resources framework was accounting for the mechanisms and processes by which models are first generated, subsequently evaluated, and then endorsed or rejected—in other words, a predictive description of why certain resources were chosen (or activated) and others were not [19]. Mechanisms such as epistemological framing [14], for example, were introduced into the resources discourse to account for the activation of specific resources in some contexts, but a growing body of research is beginning to leverage frameworks of reasoning and decision making from cognitive science to assemble a more detailed accounting of these mechanisms (see, for example, Ref. [20]). Much of this research utilizes dual-process theories of reasoning [21,22,10], which posit two types of reasoning processes in the mind. One is automatic, subconscious (intuitive), and generally fast; the other is effortful, reflective, and generally comparatively slower. These two processes are referred to as process 1 and process 2, respectively.[1] Process 1 is responsible for giving a first impression response that process 2 then follows up on (if necessary) using explicit reasoning. From a dual-process theory perspective, Heckler argued in 2011 that some incorrect responses could be explained without reference to an incorrect physics conception; instead, the response pattern could solely be attributed to lower-level cognitive factors used by process 1 to determine an answer, which may later be *justified* by process 2 using higher-level

[1] There has been an evolution of terms in the literature regarding dual-process theories. In some cases, the terms “system 1” and “system 2” are used, as in Ref. [10]; wishing to not implicate specific biological or neurological systems in dual-process theory, the terminology now preferred by Evans and Stanovich [22] is “type 1 processes” and “type 2 processes.” This paper primarily uses “process x” to refer to “type x processes.”

conceptions if specifically requested [5]. Extending the earlier analogy, students may not perceive the need to use a car at all—why use a car when walking is sufficient? As such, the origin of an incorrect response may be rooted in the lower-level cognitive factors and not in the conceptions themselves.

Heckler's argument brings into focus the need for research regarding the reasoning processes that might be impacting how students think about and answer qualitative physics questions—not only do mechanisms for model selection need to be outlined, but more specifically, the interplay between these lower-level factors and the higher-level mental constructs needs to be understood in greater detail. Along these lines, recent research has investigated several factors that affect this interplay, namely the role of processing time in questions where there are two competing dimensions (such as the slope and the height of a point on a graph) [6], the impact of perception-based bias in determining the center of mass [20], how the relative cognitive accessibility of certain ideas can influence student's performance on a wide range of tasks [7], and how the cognitive skill of mediating an intuitive, process 1 response via analytical thinking (i.e., cognitive reflection) impacts student performance on the Force Concept Inventory [9].

The presence of a salient distracting feature (SDF) [19,5,4,23,24] is another of these factors—one which has special relevance to the current investigation. Salient distracting features are features of a task that draw immediate attention away from other task features, are processed easily, and cue incorrect lines of reasoning. The salience of a feature can be operationalized by using eye tracking techniques to determine where attention is being placed. For questions in which high-salience information is irrelevant and low-salience information is relevant, it can be expected that the competition between these relevant and irrelevant features will lead to most students generating an incorrect default model based on the high salience of the irrelevant feature. Thus, the presence of a salient distracting feature represents a predictive factor that can provide insight into student answering patterns.

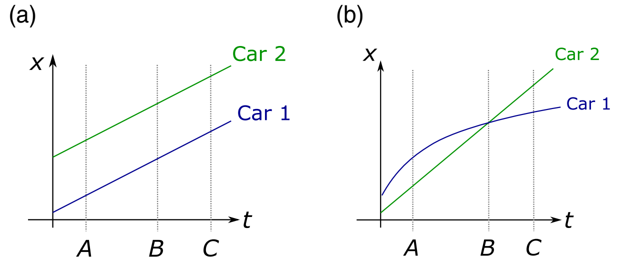


FIG. 1. Diagrams given to students as part of a study reported in Ref. [5]. The graph shown in (b) was used in the kinematics graph task (Experiments 1A and 1B) for the current work. (In the original work, the diagrams were black and white.).

Heckler demonstrated the impact of salient distracting features on physics questions by providing students with a plot of two position vs time graphs representing the motion of two cars, shown in Fig. 1 [5]. In each question, students were asked to find the time when the cars had the same speed. In one question, shown in Fig. 1(a), the two graphs were parallel lines; 90% of students chose the correct answer ("At all times"). In the other question, shown in Fig. 1(b), the two graphs intersected at time B while the slopes of the graphs were the same at a time A; 60% of students answered time A (correct), and 40% answered time B. (This difficulty with intersection points on graphs is also reported in other studies [25,26,19,5,27,28].) Notably Heckler argued that students often utilize physics concepts in defense of an incorrect time B answer cued by the salience of the intersection point [5], which highlights the interplay between low-level factors and higher-level reasoning structures.

To better understand the interplay between lower-level cognitive factors (such as SDFs) and high-level knowledge, there is a need for methodologies that separate, to the degree possible, student reasoning skills from conceptual understanding. A method for doing this, which involves paired questions, has been reported on previously [3,4]. The paired-question methodology uses a screening question that requires the student to generate a specific line of reasoning followed by a target question that effectively requires the same line of reasoning in a slightly different context. This approach then allows one to study responses from those students who answer the screening question correctly but opt for other, perhaps more salient, lines of reasoning on the target question; such students have demonstrated the ability to correctly draw upon relevant concepts in the correct line of reasoning at least once, and so their pursuit of other lines of reasoning on the target question is likely not primarily due to difficulties in conceptual understanding. This methodology is similar to "Elby pairs" [29,30], which are pairs of questions designed to elicit intuitive answers in conflict with each other; however, the task for students in Elby pairs is to resolve the conflict between their intuition and formal physics knowledge with the aim of refining intuition.

The paired question methodology was used to study a static friction task in which students are expected to reason with Newton's 2nd law to determine the magnitude of a friction force on a box that remains at rest [4]. In the screening question [see Fig. 2(a)], a single box is shown and students are told that the box remains at rest when an applied force of 30 N is acting on the box. Students are asked to compare the magnitude of the friction force to the magnitude of the applied force. The correct line of reasoning is that the box remains at rest and, by Newton's 2nd law, the net force on the box must be zero and therefore the magnitudes of the two forces must be equal to each other. Approximately 83% of students answered the screening question correctly [4]. In the

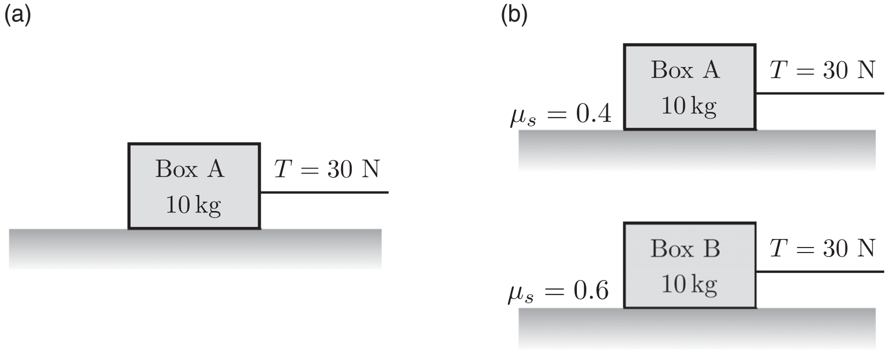


FIG. 2. Diagrams given to students for (a) the screening question and (b) the target question of the two-box friction task.

target question, students are asked to compare the forces of friction on two separate, identical boxes at rest on different surfaces while identical applied forces are exerted on both boxes [see Fig. 2(b)]. In the diagram, the coefficient of static friction for each box-surface pair is shown next to each box. From Newton's 2nd law and the observation that both boxes remain at rest, the correct conclusion is that the friction force on box A is equal to the friction force on box B. The inclusion of the two different coefficients, however, appears to elicit a common but incorrect comparison that the friction force on box A is less than the friction force on box B because the coefficient for box A is less than the coefficient for box B. Approximately 45% of students answered the target question in this particular incorrect manner, while about 65% answered correctly [4].

Of those students who answered the screening question according to the correct line of reasoning, more than 20% used the common incorrect line of reasoning on the target question [4]. This result was interpreted as a failure to engage the analytic process 2 in a productive manner. Instead, students appeared to rely on process 1 first impressions cued by the salience of the coefficients. Despite the fact that these students demonstrated the ability to step through a correct line of reasoning on the screening question, they abruptly abandoned that line of reasoning on the target question. (This interpretation was further supported by an excerpt from an interview transcript highlighting the sudden shift in reasoning approaches between the two parts.) The overall study provided further evidence that low-level cognitive influences can have an impact on the use of higher-level mental structures, but it was unclear as to how exactly this impact could be mitigated.

Low-level factors such as the salience of a specific feature can be domain general in that they impact answering patterns in predictable ways across context. For instance, the general effects of relative cognitive accessibility [7], another low-level factor related to salience, were demonstrated in the contexts of forces or friction, simple harmonic motion, kinematics, potential energy, and mass density. These low-level, domain-general influences represent mechanisms from which predictions about student answering patterns can be made; as such, understanding their impact on reasoning can provide guidance and leverage for improving student performance and reasoning skills overall. Some early efforts have been made to draw upon these mechanisms in order to improve student performance (see, for example, Ref. [8]), and the closely related investigations described in this article represent another attempt to leverage the ongoing research on cognitive mechanisms to improve student performance.

III. THEORETICAL FRAMEWORK

This work utilizes dual-process theories of reasoning as a theoretical framework. These theories propose two separate processes in the mind by which reasoning and decision making occur. Process 1 is primarily at play in decisions that rely on automated responses such as how to manipulate a steering wheel to keep a car in the center of a lane or judging someone's emotions from a glance at that person's face. Process 1 guides much of adult decision-making throughout the course of a day because it is optimized to reduce cognitive load and free up working memory for more important tasks (i.e., humans tend to be cognitive misers). When there is a reason to expend effort, process 2 comes into play recruiting working memory to run simulations, test hypotheses, or execute an algorithm. This process is helpful with problems such as long division or deducing a result from first principles.

Among the general theories of reasoning that fall under the umbrella of dual-process theories, we have found the heuristic-analytic theory [21] to be particularly helpful in analyzing student responses to our physics tasks. The heuristic-analytic theory of reasoning, shown diagrammatically in Fig. 3, describes three main mechanisms (formulated as principles) by which mental models are generated, evaluated, selected, and/or abandoned. These principles are *the relevance principle, the singularity principle*, and *the satisficing principle* [21].

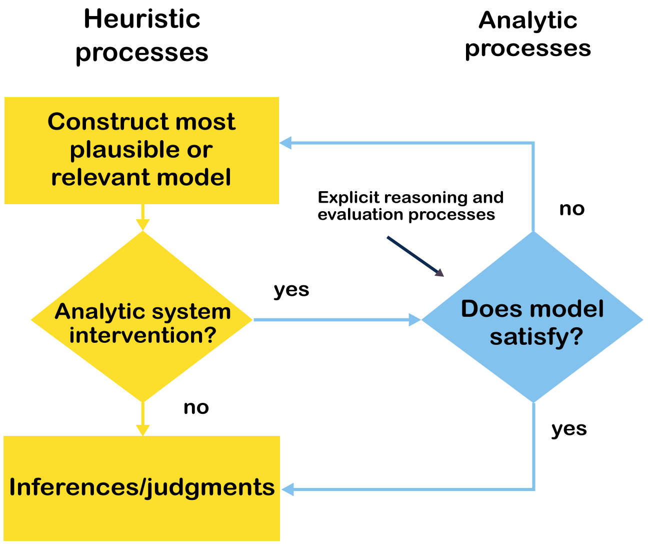

Heuristic processes
Analytic processes
Construct most plausible or relevant model
Explicit reasoning and evaluation processes
no
yes
Analytic system intervention?
Does model satisfy?
no
yes
Inferences/judgments

FIG. 3. Diagram showing the separate roles of the heuristic (type 1) and analytic (type 2) processes, taken from Ref. [21].

In the heuristic-analytic theory, process 1, the *heuristic process*, is responsible for generating a mental model to serve as an entry point into any reasoning path. In this context, a mental model is a mental representation of the structure or relationships between given entities. According to the *singularity principle*, only one mental model is considered at a time. The *relevance principle* states that this first-available model (or *default model*) is chosen based on the perceived relevance of the model to the current task—which in turn is informed by task features (e.g., contextual cues), epistemological frames, and prior knowledge.

Another key aspect of the default model is that it is accompanied by a value judgment about how plausible the model is, referred to as a *feeling of rightness* [31]. This is a measure of how confident a reasoner is that the model is appropriate for the task at hand. If the feeling of rightness is strong, a reasoner may proceed to make a judgment directly from the default model; if the feeling of rightness is sufficiently weak, process 2, the *analytic process*, is engaged. Some individuals have a general disposition toward reflective thinking [32] or *cognitive reflection* [33] (i.e., they have a tendency to mediate process 1 output by reasoning more analytically), and therefore develop a habit of mind to engage process 2 in order to scrutinize the default model when there is a sense of cognitive strain [10].

The engagement of the analytic process is referred to as an *analytic intervention*. According to the *satisficing principle*, process 2 is primarily concerned with ascertaining whether or not the default model is satisfactory for the task at hand. Because process 2 must decide whether or not to endorse the default model, it is necessarily influenced by the feeling of rightness in the default model, and reasoning biases such as confirmation bias [34] can also enter into a reasoner's thinking and decision making. Explicit process 2 reasoning relies on *mindware*, a term coined by Perkins and further extended by Stanovich to refer to the collection of "rules, knowledge, procedures, and strategies that a person can retrieve from memory in order to aid decision making and problem solving." [35]. In accordance with the singularity principle, alternate models will be explored only if the default model is found to be unsatisfactory by the analytic process, at which point the flow chart depicted in Fig. 3 is repeated.

This theory has implications for student behavior when responding to a qualitative inferential reasoning task in physics (by which we mean a problem which requires students to step through a series of inferences using physics concepts to arrive at a final conclusion). Since reasoning occurs using one model at a time (the singularity principle) and the process by which a model can be identified as unsatisfactory has to be activated (e.g., by a decreased feeling of rightness in the initial model), incorrect physics models cannot be abandoned in the moment without sufficient evidence. However, while the intervention of the analytic process is necessary, it is not sufficient for abandoning an incorrect default model; a student must also possess relevant *mindware* for solving the problem correctly. Even if the default model is not accompanied by a strong feeling of rightness, it will still be used to make judgments in the absence of the mindware necessary to generate a satisfactory correct model. As a result, a productive analytic intervention requires both that the analytic intervention be triggered in a meaningful way *and* that the student possesses the relevant mindware to rule out the default model and make progress with a correct model.

Thus, the theoretical framework described above leads us to the following working hypothesis:

> *An analytic intervention that results in abandoning an incorrect default model is more likely to occur if and only if (1) students are presented with information that refutes the default model as opposed to information that promotes a correct model, and also (2) students possess the mindware necessary for replacing the default model with a correct model.*

Because process 2 works to refute the default model *before* alternate models are considered, information that supports alternate models (e.g., a correct model) is likely to either lie unexamined or be used in association with the default model, even if that information is inconsistent with the default model. Thus, a corollary to the working hypothesis of the paper is the following:

> *For students with an incorrect default model, information in support of a correct model is likely to be incorporated into reasoning that supports the default model instead of promoting the abandonment of the default model.*

Together, this working hypothesis and corollary provide the theoretical basis for the experiments described in this article.

In our investigation, we used reasoning chain construction tasks as a venue in which to explore the extent to which dual-process theories of reasoning (as articulated in the working hypothesis and corollary) can successfully predict student reasoning and answering patterns on certain physics questions. In particular, our investigation focused on the following research questions, which guided our methodology and experimental design:

RQ1. How, if at all, does providing students with correct statements in support of a correct model impact student answering patterns on a physics question containing one or more salient distracting features?

RQ2. How, if at all, does providing students with a statement that refutes an incorrect default model impact student answering patterns on a physics question containing one or more salient distracting features?

RQ3. To what extent is the impact of providing students with a statement that refutes an incorrect default model mediated by the presence of relevant mindware?

From our working hypothesis, we predicted that providing students with correct statements would not impact student answering patterns (RQ1). We expected, however, that providing a statement that refutes an incorrect default model *would* impact student answering patterns generally (RQ2), and furthermore that this impact would be limited to those who demonstrate relevant mindware (RQ3).

IV. METHODOLOGY AND EXPERIMENTAL DESIGN

In this section, we present a new methodology that helps disentangle reasoning approaches from conceptual understanding and foregrounds domain-general reasoning phenomena. We then describe two experiments that highlight the affordances of this methodology in probing the extent to which dual-process theories of reasoning can explain student reasoning in physics.

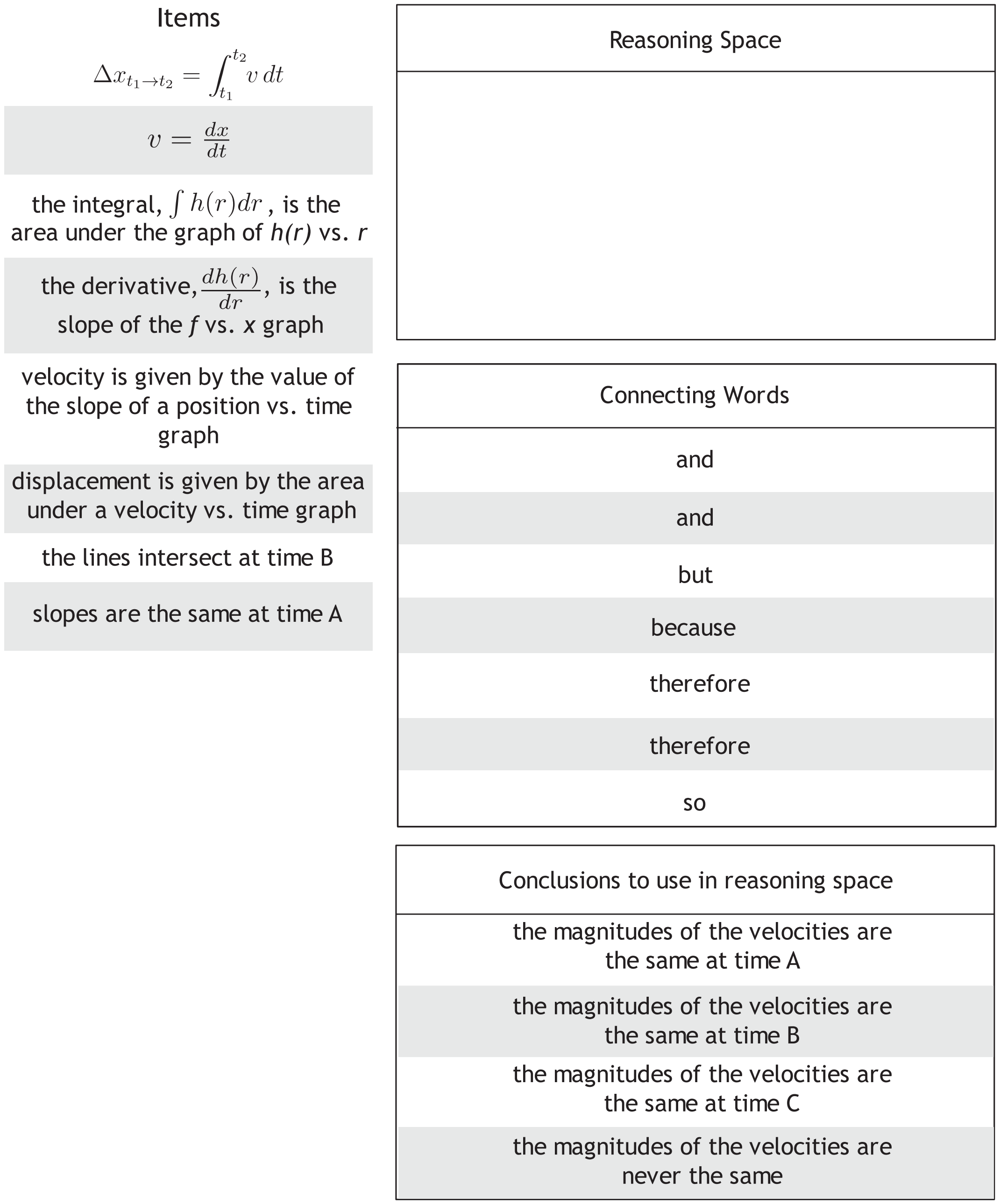
A. A new methodology: The reasoning chain construction task

The methodology we developed and employed centers around a *reasoning chain construction task*, or simply a *chaining task*, which allows students to focus on arranging statements of conceptual knowledge and observations about the physical context into a logical progression of inferences. To accomplish this, we (i) provide the student with a list of reasoning elements; (ii) indicate that all of the statements within these elements are true and correct; and (iii) ask the student to construct a solution to a physics problem by selecting elements from the list, ordering them, and, as needed, incorporating provided connecting words ("and," "so," "because," and "but"). The reasoning elements primarily consist of observations about the problem setup, statements of physical principles, and qualitative comparisons of quantities relevant to the problem, all of which are true. Everything the student needs to produce a complete chain of reasoning is present in the elements; the student's task is then to pick from given conceptual pieces and assemble a reasoning chain. For this investigation, we focused on tasks requiring only a few steps.

Reasoning chain construction tasks have primarily been implemented online using the Qualtrics survey platform [36], using the "Pick/Group/Rank" question format. This online format is illustrated in the context of a graph task and is shown in Fig. 4. Reasoning elements from the "Items" column, connecting words, and final conclusions can all be dragged and dropped into the "Reasoning Space" box; the box increases in size vertically as elements are added.

These tasks were administered on special participation-based homework assignments and exam reviews for students enrolled in an introductory calculus-based physics sequence, along with other questions relevant to the course but not directly related to the content targeted by the research task. These assignments counted for participation credit or extra credit (i.e., students received full credit regardless of the correctness of their responses), and differed from the standard online homework assigned in the course. Participation rates for these assignments typically ranged from approximately 45% of students enrolled to above 95%, with an average participation rate of about 70% across the experiments. Given that these special ungraded assignments were for participation credit and typically intended for individual exam review and preparation, we suspect that student collaboration was relatively uncommon. (For this reason, we did not explore the potential impact of even a small amount of student collaboration on this study's tasks.) In all cases, the tasks were administered at a research university in New England after relevant lecture, laboratory, and small-group recitation instruction (i.e., after lecture, lab, recitation instruction as well as homework questions on the topic, but before exam coverage of the topic). Research-based materials from *Tutorials in Introductory Physics* [37] were used in the recitation sections.

The reasoning elements provided to the students were informed by previously obtained student responses to open-ended, free-response versions of a given task. Some elements were productive to the correct line of reasoning, and some were not. Among the unproductive elements were those which, while true, were useful primarily in constructing the common incorrect line of reasoning. In addition, the extent to which students' final responses contained unproductive elements not associated with the common incorrect line of reasoning helped us gauge the likelihood that students were not taking the task seriously and rather were simply inserting elements at random. In



Items

$\Delta x_{t_1\rightarrow t_2} = \int_{t_1}^{t_2} v\,dt$

$v = \frac{dx}{dt}$

the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*

the derivative, $\frac{dh(r)}{dr}$, is the slope of the *f* vs. *x* graph

velocity is given by the value of the slope of a position vs. time graph

displacement is given by the area under a velocity vs. time graph

the lines intersect at time B

slopes are the same at time A

Reasoning Space

Connecting Words

and

and

but

because

therefore

therefore

so

Conclusions to use in reasoning space

the magnitudes of the velocities are the same at time A

the magnitudes of the velocities are the same at time B

the magnitudes of the velocities are the same at time C

the magnitudes of the velocities are never the same

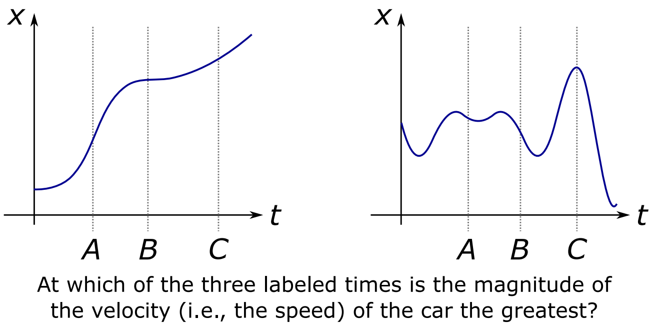
FIG. 4. Example of how a chaining task appears to students online via the Qualtrics platform. Note that the diagram, shown in Fig 1(b), and the question prompt from the kinematics graph task were also provided to students.

practice, such responses were very rare typically accounting for less than 2% of total responses on a given task. Three blank elements labeled “Custom:” were provided, with instructions that students could use the text box attached to the custom element to create their own reasoning element(s) if students felt they wanted to add something absent from the given reasoning elements.

In the next sections, we briefly describe each experiment in order to provide an overarching view of the experimental design used in this investigation. Details as well as results are discussed in Secs. V and VI.

B. Experiment 1A and 1B: Examining the impact of statements that support a correct model

Experiments 1A and 1B were designed to test the hypothesis that the inclusion of information that supports a correct model is not enough to help students disengage from an incorrect default model. In testing this hypothesis,



At which of the three labeled times is the magnitude of the velocity (i.e., the speed) of the car the greatest?

FIG. 5. Screening questions used to gauge ability to determine the magnitude of velocity from a position vs time graph. Both graphs were presented to the student along with the prompt shown.

both experiments directly addressed RQ1 (impact of statements supporting a correct model on student answering patterns). These experiments also tested the corollary that if a default model is not abandoned, the information would instead be used to justify that model—even if that information appears to an expert to be inconsistent with the default model.

For experiment 1A, we cast the kinematics graph task (KGT) from Ref. [5] [see Fig. 1(b)] as a reasoning chain construction task. We also developed two screening questions that were meant to gauge whether a student possessed the relevant mindware required to determine the magnitude of an object's velocity from a position vs time graph. These two screening questions are shown in Fig. 5.

In our experiments, students were randomly placed via Qualtrics in either a treatment or control condition. In the treatment condition, students were given the chaining task version of the kinematics graph task; in the control condition, students were given the kinematics graph task in a more standard multiple-choice format followed by a prompt to explain the reasoning they used to arrive at an answer. All students were given the screening questions in a multiple choice with explanation format (a question format commonly used in their physics courses). Since we wanted to ensure that the act of completing the screening questions would not impact student performance on the kinematics graph task (e.g., by priming student thinking), the screening questions were placed after the kinematics graph task and separated from it by several questions on unrelated topics.

Experiment 1B tested the domain-general nature of the salient distracting feature and was meant to further examine the hypothesis that information that promotes a correct model would not readily cause students to abandon the default model. In experiment 1B, three tasks isomorphic to the kinematics graph task were devised in the contexts of mechanical potential energy, electric potential, and magnetic flux (similar to the isomorphs in Ref. [6]). Each task used the same plot with identical intersecting graphs, and the wording in the plots was kept as parallel as possible while reflecting the new contexts. Additionally, the reasoning elements provided on the kinematics graph task were altered slightly to reflect the new context but were otherwise parallel and isomorphic in structure to those on the kinematics task. The problem statements and reasoning elements for these three tasks are provided in the Appendix. Isomorphic screening questions were similarly constructed.

The design for experiment 1B was the same as that for experiment 1A: students were randomly placed in a treatment condition (chaining task) or a control condition (multiple choice with explanation). In each case, the screening questions were placed after the graph task and separated from it by multiple questions on unrelated topics. Given that the four graph tasks were administered across a single academic year, most students who completed the introductory calculus-based sequence would have seen and completed multiple, and possibly all four, tasks.

C. Experiment 2A and 2B: Examining the impact of a statement that refutes the incorrect default model

Experiment 2A was designed to test the main working hypothesis that, for students with an incorrect default model, providing information that refutes the default model will be more effective in supporting productive analytical engagement than information that supports a correct model. This experiment therefore addressed RQ2. In this experiment, we cast the two-box friction task from Ref. [4] [see Fig. 2(b)] as a reasoning chain construction task and randomly assigned the students into treatment and control conditions. Both conditions utilized the chaining format version of the friction task, but in the treatment condition, a single element was added to the list of reasoning elements provided to the student. This element indicated that "the coefficient of static friction is not relevant to this problem" and was designed to call into question student satisfaction with the common, incorrect default model, thereby promoting cognitive reflection and productive analytical engagement.

In experiment 2B, we administered the screening question (in multiple choice with explanation format) reported in Ref. [4] and shown in Fig. 2(a) prior to the chaining task. In contrast to experiments 1A and 1B, the screening question was placed immediately prior to the chaining task to replicate earlier studies using this screening or target pair [4]. The screening question allowed us to test the hypothesis that presence of relevant mindware is required for a productive engagement of the analytic process resulting in the selection of a correct alternate model. This experiment therefore addressed RQ3. In the experiment, we operationalized the possession of correct mindware as the demonstrated ability to answer the screening question correctly with correct reasoning.

D. Statistical analysis methods

When comparing control and treatment conditions throughout this investigation, we performed Fisher's exact tests to compare answer choice distributions, and we ascertained the statistical significance and effect size of our experimental results by examining the associated p value and Cramer's V for each test. Fisher's exact test determines whether the answer choice distribution from a control condition is different enough from an answer choice distribution from a treatment condition that the difference is unlikely to be due to chance alone and is an appropriate test because the variable under study (answer choice) is *categorical*. Fisher's exact test is superior to the chi-square test when N values are lower (such as in some of the experiments in this article) because it provides an exact measure of a probability rather than an approximation. Specifically, in this paper, Fisher's exact test tests the hypothesis that observed frequencies of answer choices in one population come from the same distribution as observed frequencies of answer choices in another population. If the test returns a p value below 0.05, there is a less than 5% chance that the distributions are the same. This threshold p value is assumed to be an indicator below which the hypothesis is false—the two distributions are different, implying that the two distributions come from independent populations. In the context of this work, finding that two populations (i.e., control and treatment) are different is interpreted to mean that the treatment condition altered the state of the students in that condition such that they are now different than the students in the control condition—at least for the amount of time it took to participate in the experiment(s). This last point is important as the work in this study doesn't claim to produce long-lasting effects, but rather in-the-moment changes to the dynamic process of reasoning.

Our questions have four answer choices: a correct answer, a common incorrect answer, and two *other* incorrect answers that are not as commonly selected. In this article, we report the full distribution of answer choices, but for the Fisher's exact tests we collapsed the distribution to just *correct, common incorrect*, and *other*. We did this because the common incorrect answers associated with the salient distracting features are the focus of this study, so separating these incorrect answers from the other incorrect answers is of value.

In addition to providing p values, we also report the effect size (via Cramer's V) for each test. While Cramer's V is typically used to measure the effect size of a chi-square test, it is also appropriate to use for a Fisher's exact test because its computation is not reliant on any specific hypothesis testing, but rather on the distributions themselves. The effect size is an indication of the magnitude of difference found between the two distributions. What constitutes a large effect size depends on degrees of freedom (i.e., the number of answer choices minus one multiplied by the number of conditions minus 1). For the tests in this manuscript, there were 2 degrees of freedom, so an effect size less than 0.07 is considered negligible, an effect size between 0.07 and 0.2 is classified as small, an effect size between 0.2 and 0.35 is classified as medium, and one of 0.35 or higher is classified as large. We also employ residual analysis to ascertain information about the nature of the difference between the two distributions. Residuals are a measure of the deviation of counts in a category from the expected counts if the distribution of counts in each condition were the same. Residuals can therefore provide some evidence regarding which specific answer choices are different when comparing control to treatment. A standardized residual greater than 2 (indicating more than 2 standard deviations from the expected count) is the typical benchmark for a noteworthy deviation.

V. EXPERIMENTS 1A AND 1B: EXAMINING THE IMPACT ON STUDENT ANSWERING PATTERNS OF STATEMENTS THAT SUPPORT A CORRECT MODEL IN THE CONTEXT OF GRAPH TASKS

Experiments 1A and 1B were designed to test the first part of the working hypothesis—namely, that information that supports a correct model is not enough to help students disengage from an incorrect default model. In testing this hypothesis, we directly addressed RQ1. These experiments also tested the corollary that if a default model is not abandoned, the information would instead be used to justify the default model, even if that information appears to an expert to be inconsistent with the default model.

A. Experiment 1A: Using the kinematics graph task to examine the impact of statements that support a correct model

In this section, we provide an overview of experiment 1A, describe our predictions, and discuss the results from the task and accompanying screening questions.

1. Description of experiment 1A

In experiment 1A, we cast the kinematics graph task [KGT, shown in Fig. 1(b)] as a chaining task, with the reasoning elements shown in Table I. Four of these elements (bold text in Table I) can be productive to the correct line of reasoning. These four elements have an implicit logical structure. While at first glance, it may appear that the elements "$v = dx/dt$," "*the derivative*, $dh(r)/dr$, *at a specific point is the slope of the tangent line of the* $h(r)$ *vs r graph at that point*," and "*velocity is given by the value of the slope of a position vs time graph*" are equivalent and largely interchangeable statements, they actually constitute a logical argument justifying why the slope is the velocity: the two elements "$v = dx/dt$" and "*the derivative[…] is the slope…*" combine to imply the third element. (In this paper these three elements are collectively called the *velocity triad*.) We refer to the element "*velocity is given by the value of the slope of a position vs time graph*" as a derived heuristic because it

TABLE I. Reasoning elements provided to the students on the kinematics graph task. Elements productive to the correct line of reasoning (i.e., elements that support a correct model) are bolded.

$\Delta \mathrm{x}_{\mathrm{t}_1 \to \mathrm{t}_2} = \int_{t_1}^{t_2} v(t)dt$
$\boldsymbol{v = dx/dt}$
the integral, $\int h(r)dr$, is the area under the graph of $h(r)$ vs r
the derivative, $\boldsymbol{dh(r)/dr}$, at a specific point is the slope of the tangent line of the $\boldsymbol{h(r)}$ vs $\boldsymbol{r}$ graph at that point
velocity is given by the value of the slope of a position vs time graph
displacement is given by the area under a velocity vs time graph
the lines intersect at time B
the slopes are the same at time A
the magnitudes of the velocities are the same at time A
the magnitudes of the velocities are the same at time B
the magnitudes of the velocities are the same at time C
the magnitudes of the velocities are never the same

represents a chunked knowledge piece [38] that is derived from two independent principles. While it would be acceptable to many instructors if students were to simply use the "slope is velocity" heuristic, all three elements are needed to provide a logically sound argument. Their inclusion provided an opportunity to gain additional insight into the extent to which students reason on the basis of derived heuristics vs foundational principles.

For this experiment, a between-student design was employed with the treatment condition corresponding to the chaining version of the graph task, and the control condition corresponding to a multiple choice with explanation version of the graph task. Two screening questions were also administered in multiple choice with explanation format. The two screening questions, shown in Fig. 5, ask students to determine the time at which the magnitude of velocity was the greatest. The screening questions contain distractors that tend to elicit slope or height confusion and difficulties in interpreting a negative vs a positive slope. We considered correct responses to both screening questions to serve as an indicator of the presence of mindware relevant for successful reasoning on the graph task.

2. Predictions drawn from working hypothesis

Regarding RQ1 (impact of statements supporting a correct model on student answering patterns), we hypothesized that students will not abandon a default model unless there is sufficient reason to question their satisfaction with that model; and as a corollary, that information supporting a correct model would be recruited to defend the default model rather than to abandon it. This hypothesis led to specific predictions for student behavior in experiment 1A.

The high-salience intersection point typically results in many students embracing a default, intersection-cued model, leading to an answer of time B (the time at which the two graphs intersect). Since no reasoning elements were provided to the students that explicitly refute this incorrect default model, we predicted that *explicit inclusion of reasoning elements associated with a correct line of reasoning will not greatly impact student answering patterns on the task since it will not preclude students from endorsing the incorrect default model* (prediction 1).

Because the high salience of the intersection point affects process 1 reasoning and is not necessarily connected with models based in physics content, we would expect the default model to be associated with the intersection regardless of whether or not someone possessed mindware relevant to obtaining the velocity from a position vs time graph. Because this mindware will not likely be employed in the absence of dissatisfaction with the default model, we also predicted that *explicit inclusion of reasoning elements associated with a correct line of reasoning will not greatly impact student answering patterns even among those students who correctly answer both screening questions* (prediction 2).

Finally, because of the satisficing principle, if the default model is not abandoned, process 2 will likely utilize formal reasoning to justify the default model—even if that reasoning is logically flawed or inconsistent with other reasoning provided by the student elsewhere. Thus, we predicted that *elements productive to the correct line of reasoning would likely be incorporated into the reasoning chains in support of the incorrect default model* (prediction 3).

3. Analysis of answer choice distributions and discussion

Student answer choice data from the chaining version of the kinematics graph task (treatment) from a single semester are shown in Table II, along with data from the multiple choice with explanation version of the task (control) administered the same semester. As can be seen in Table II, there is a statistically significant but small difference in the answer distribution between treatment and control ($p = 0.025$, $V = 0.16$). On inspection of the residuals, the percentage of common incorrect answers remained the same—the residual for time B (the highly salient intersection point) was -0.3, whereas the residual for time A was $+2.2$ and the residual for time C or never

TABLE II. Student answer distribution data from two versions (control and treatment) of the kinematics graph task (KGT) administered in experiment 1A. The task itself is shown in Fig. 1(b). There is a small difference in the answer distribution on the chaining format in comparison with the multiple choice with explanation format ($p = 0.025$, $V = 0.16$). The correct answer choice is in boldface for reference.

	Percentage of total responses	
	KGT control (MC with explanation)	KGT treatment (chaining format)
N	158	149
Time A (correct)	**44%**	**57%**
Time B (intersection, common incorrect)	30%	29%
Time C	1%	0%
Never	24%	14%

was −2.5. This result suggests that the presence of correct, relevant reasoning elements alone was not enough to reduce the number of answers focused on the intersection.[2]

From either a misconceptions or resources perspective, this result may be explainable but is hard to predict. For instance, it has been argued that students who select the intersection in the KGT lack a conceptual understanding of velocity, are drawing upon incorrect ideas about velocity, or are cued to construct models around the "same is same" resource in which the height becomes relevant. By providing the relevant, correct conceptual elements, one might predict that the prevalence of correct answers should increase considerably because students may now draw upon these elements, which might help them refine their understanding of velocity, address an incorrect concept, or redirect the "same is same" resource to the alternate cue "the slopes are the same at time A." However, because there are not well-defined mechanisms for what specific knowledge is constructed or accessed in the moment, no firm prediction can immediately be made.

Dual-process theories of reasoning, however, make a firm prediction because they give more definition to the control mechanisms by which models are chosen for consideration as well as the conditions under which they would be abandoned in favor of alternate models. In this case, an incorrect model based on the intersection point drew some students to the time B answer. In order for students to switch away from this default answer, an analytic intervention would need to be triggered (i.e., a productive engagement of process 2 associated with cognitive reflection), resulting in a loss of confidence in this answer. However, the default mental model from process 1 is the entryway into any path of reasoning and thus impacts the subsequent reasoning process. Since humans tend to be relatively poor at coming up with and exploring counterarguments and often seek to rationalize the default model (resulting in reasoning biases), the analytic process is more likely to be engaged in a somewhat superficial manner and to identify physicslike justifications for the original model rather than to systematically rule out that model and ultimately arrive at a different answer. The presence of correct information alone would therefore not be expected, for many students, to produce the level of dissatisfaction required to prompt an in-depth scrutinization of the incorrect default model and subsequent exploration of alternate models (consistent with our articulation of prediction 1).

4. Analysis of answer distributions and discussion: Screening questions

According to prediction 2, we would expect that even among those students who demonstrate the mindware needed to obtain the magnitude of velocity from a position vs time graph on the screening questions, the treatment condition would likely not yield a large difference in answer distribution on the KGT despite increased access to relevant conceptual statements. We would thus expect that the intersection point would still be a prevalent incorrect answer among those who have elsewhere demonstrated the requisite mindware.

Overall, student performance on the screening questions (see Fig. 5) was rather strong. Ninety-six percent of participants ($N = 307$) correctly answered screening question 1, 83% of students correctly answered screening question 2, and 82% correctly answered both. It is worth noting that the screening questions included distractors consistent with slope-height confusion. In both questions, time C had the greatest height. This answer was not prevalent in screening question 1, but it accounted for 17% of student responses to screening question 2. It is surmised that the particular shape of the graph contributed to this difference in prevalence of responses indicative of slope-height confusion, with the sharpness of the curve at time C in question 2 possibly being more salient than the smooth curve at time C in question 1. This speculated difference in salience is also consistent with previous research on salient distracting features in graphs [19].

For those students who answered both screening questions correctly, the observed difference in answer distribution (*correct, common incorrect, and other*) from control to treatment was statistically significant with a small effect size ($p = 0.046$, $V = 0.16$). Additionally, 22% of students in the treatment condition who answered both screening questions correctly (thereby demonstrating the requisite

[2] In another study, we used the affordances of the chaining format to track the dynamics of student reasoning chains *as they were being constructed* and found that if a student switched their answer during the course of constructing a chain, it was away from time B, not toward it. This suggests that the difference in the end results shown in Table II is due to switching from "Never" to "Time A" rather than more complex dynamics.

mindware, $N = 122$) ultimately chose time B on the KGT, which corresponds to the intersection point. Residual analysis shows that the time A answer choice has a larger than expected number of counts in the treatment condition (+2.39) but that the time B answer choice did not show a large difference from deviation (residual of −1.13); instead, the larger residual came from the time C or never choices, which had a residual of −1.77. This suggests that the increase in the time A answer choice compared to the control group primarily came from those who answered either "time C" or "never." Taken together, these results are consistent with the expectations expressed in prediction 2, namely, that even among those students who demonstrated the mindware needed to obtain the magnitude of velocity from a position vs time graph on the screening questions, access to the relevant concepts and information would not yield a large change in the answer distributions on the KGT. The results illustrate how a compelling model arising from type-1 processing may hinder student ability to access and leverage the relevant knowledge they possess; indeed, almost one-quarter of students who demonstrated the requisite mindware arrived at an answer that was inconsistent with that mindware—even when relevant information was explicitly provided.

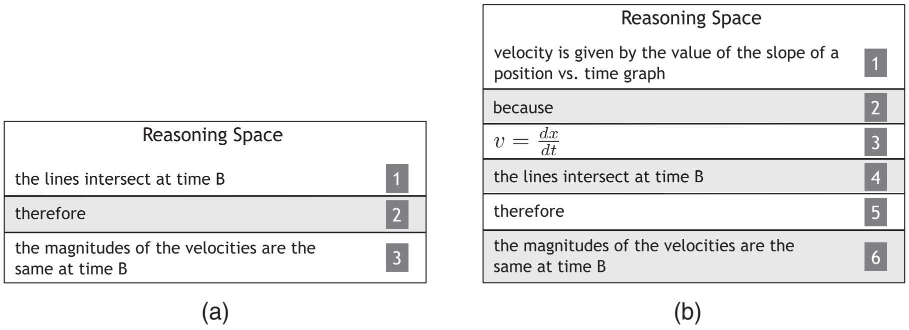
5. Analysis of incorrect reasoning chains on the kinematics graph task

The chaining format affords students an opportunity to employ reasoning elements that they might otherwise not consider using. According to the dual-process framework, we predicted that such reasoning elements would likely also be used in conjunction with the incorrect default answer put forward by process 1, even if the element itself was inconsistent with the default answer (prediction 3). We analyzed in detail the reasoning chains constructed by students in support of the common incorrect answer.

Among those students who selected time B ($N = 44$) as their answer, a substantial number (44%) constructed chains that only included both the reasoning element "the lines intersect at time B" and the conclusion element "the magnitudes of the velocities are the same at time B." We categorized such chains as being the canonical incorrect response, illustrated in Fig. 6(a), as the chains were similar to the kinds of incorrect free-response justifications observed on this task. Indeed, these chains simply articulated the salient distracting feature (the intersection) along with the incorrect answer, with no reference to any physics concepts.

A larger number of students (56%, $N = 44$) also used productive elements inconsistent with their answer in supporting the common incorrect response. We classified such responses in the *struggle response* category as they seemed indicative of an apparent tension between correct mindware and an incorrect default model. An example of such a response is shown in Fig. 6(b). The first three elements, "velocity is given by the value of the slope of a position vs time graph, because, and $v = dx/dt$" are logically connected in a way that, to an expert, suggests an understanding of the underlying physics. Indeed, this student explicitly endorsed correct conceptual information before abruptly shifting to the incorrect answer associated with the salient distracting feature.

To study this phenomenon in greater detail, criteria were developed to gauge the extent to which students who both chose the intersection (time B) and endorsed productive elements were demonstrating understanding of the underlying physics. The most rigorous criterion required the student to use two or more of the three elements that comprise the velocity triad described in Sec. V A 1, as in Fig. 6(b). In all cases in which a student satisfied this criterion, it was clear that the student was linking the elements together logically. Of those students who answered time B, 7% met this requirement. The second, more relaxed criterion contends that any student who uses at least one of the three elements (without using any irrelevant elements) is endorsing correct conceptual information. This is appropriate given that the derived heuristic element, "velocity is slope," is commonly the only element used in supporting a correct answer. It also represents an idea that is likely to be highly accessible to a student due to its ubiquity during classroom instruction on kinematics. Use of this criterion raised the proportion of students who


Reasoning Space

the lines intersect at time B 1

therefore 2

the magnitudes of the velocities are the same at time B 3

(a)

Reasoning Space

velocity is given by the value of the slope of a position vs. time graph 1

because 2

$v = \frac{dx}{dt}$ 3

the lines intersect at time B 4

therefore 5

the magnitudes of the velocities are the same at time B 6

(b)

FIG. 6. (a) A canonical incorrect response in which a student solely justified the answer on the basis of the observation that the lines intersect at time B. (b) A response in which a student endorses information more closely associated with the correct line of reasoning in the process of justifying the common incorrect answer.

both chose time B and certified correct information to over 55%. We were thus able to generate both lower (7%) and upper (56%) bounds on the extent to which students who chose the intersection were demonstrating some level of understanding of the underlying physics in their chains.

These results indicate a sort of tension between relevant mindware and an intuitive answer generated by process 1. Our prediction was that some students who chose time B, when confronted with improved access to knowledge relevant to the correct line of reasoning, would incorporate this contradictory knowledge into a reasoning chain in support of the common incorrect answer. This prediction proved to be correct, with up to 56% of students ($N = 44$) who chose the common incorrect answer using elements in their chain that represented reasoning that, to an expert, is inconsistent with the answer itself.

B. Experiment 1B: Using multiple contexts to examine the impact of statements that support a correct model on student answering patterns

Experiment 1B extends the results of experiment 1A across three additional contexts: potential energy, electric potential, and magnetic flux.

1. Description of experiment 1B

Based on dual-process theories of reasoning, the intersection point on a graph in contexts outside of kinematics should result in prevalent incorrect responses based on the same default judgment cued by the intersection point. Indeed, even in contexts outside of kinematics, process 1 will rely on the salient features of a task when selecting an initial model. Heckler and Scaife used math graphs, kinematics graphs, and graphs of electric potential to demonstrate that processing time had an effect on answer patterns for questions regarding the slope of a graph independent of context [6]. While context and content mediate the effects of domain-general factors, these factors are still at play. For instance, in Heckler and Scaife's work, the effects of processing time were less pronounced in more familiar contexts but were still present [6]. Likewise, the working hypothesis of this paper (i.e., that access to relevant conceptual information would not be sufficient for most students to abandon an incorrect default model) should be operative regardless of specific physics content.

To test this hypothesis, three additional chaining tasks were devised. These tasks were structurally parallel to the kinematics graph task and were in the contexts of potential energy, electric potential, and magnetic flux. For each context, the correct line of reasoning relies on an understanding that the desired quantity can be obtained from the derivative of the graphed quantity, and thus the slopes of the graphs at the point of interest should be compared. We constructed screening questions that would indicate the extent to which the students possessed the ability to determine the desired quantity from slope in the absence of the intersection. The reasoning elements provided to the student in each task were modified to fit the context but remained isomorphic in their structure. All graph task prompts, the provided reasoning elements, and the associated screening questions are included in the Appendix.

All tasks were administered after relevant course instruction (i.e., after lecture, lab, recitation instruction as well as homework questions on the topic, but before exam coverage of the topic). Given the contexts associated with these isomorphic tasks, data were collected in both semesters of the on-sequence calculus-based introductory physics sequence. The experimental design was the same as that used with the kinematics graph task; a between-student design was employed with the treatment condition corresponding to the chaining version of the graph task, and the control condition corresponding to a multiple choice with explanation version of the graph task.

2. Predictions drawn from working hypothesis

Given the similarity in experimental design, we expected all three predictions (see Sec. V A 2) made for experiment 1A to hold for the isomorphic graph tasks in experiment 1B as well. The three additional graph tasks, however, could help us generalize our results from experiment 1A.

3. Analysis of answer distributions and discussion

In this section, we first examine and discuss the general performance on all graph tasks and then consider the results

TABLE III. Answer distribution comparison between control (multiple choice with explanation) and treatment (chaining format) for each graph task in experiment 1B. The tasks themselves are shown in the Appendix. †Data collected from the previous year for magnetic flux task. See Sec. V B 3. The correct answer choice is in boldface for reference.

	Kinematics		Potential energy		Electric potential		Magnetic flux†	
	Control	Treatment	Control	Treatment	Control	Treatment	Control	Treatment
N	158	149	80	76	121	97	83	88
Time A (correct)	**44%**	**57%**	**38%**	**43%**	**44%**	**73%**	**59%**	**66%**
Time B (intersection)	30%	29%	58%	51%	45%	21%	40%	28%
Time C	1%	0%	0%	1%	3%	1%	0%	5%
Never	24%	14%	5%	4%	8%	5%	1%	1%
(p, V)	(0.025, 0.16)		(0.75, 0.06)		(0.001, 0.30)		(0.11, 0.16)	

from the screening questions. The results from all four isomorphic graph tasks are summarized in Table III.

There is little or no statistically significant difference in answer distributions on the chaining version in comparison to that on the multiple choice with explanation version for two of the three new graph tasks. Residual analysis also shows that for the potential energy task and the magnetic flux task, the time B answer does not change considerably. (The residuals for the time B answer choice were −0.76 for the potential energy task and −1.57 for the magnetic flux task, whereas the time B residual was −3.72 for the electric potential task.) These results, combined with those from the KGT, suggest that providing increased access to relevant physics concepts does not greatly reduce the prevalence of intersection answers. The electric potential graph task was the one exception, as it exhibited a positive, medium effect-size difference on the treatment version in comparison to the control version. We discuss this discrepancy later in this section.

As a side note, it was not possible to collect truly analogous multiple choice with explanation data for the magnetic flux task given a different experiment we were conducting as part of our broader investigation. As such, data collected the previous year from both versions (treatment and control) of the isomorphic flux graph task are included in Table III. However, the results are similar to those collected for the flux task administered in the same year as the other three tasks. The results for the magnetic flux tasks shown in Table IV, however, are drawn from data collected in the same academic year as the other three tasks.

Chaining format results for those students who answered both screening questions correctly, thereby demonstrating that they possess the requisite mindware, are shown in Table IV. The intersection point remains a common incorrect answer in all three additional tasks, with around 24% of the students picking time B across all four tasks.

On the electric potential graph task, however, inclusion of correct reasoning elements does result in a difference in answer distributions of medium effect size. The impact of the reasoning elements in this case appears to be content specific (i.e., somehow related to the topic of electric potential), but we are unsure of the particular cause. Electric potential and electric field are easily confused [39,40], perhaps in a way similar to that of position and velocity. However, the electric potential-electric field relationship does not share the same degree of intuitive knowledge (i.e., “folk physics”) or firsthand experience as that between position and velocity. One might speculate, therefore, that students are more likely to turn to the provided reasoning elements in order to figure out in the moment how to differentiate electric potential and electric field and subsequently establish an appropriate relationship between these two easily confused concepts, leading to noticeably stronger performance. Still, despite the observed difference associated with the treatment condition, the salience of the intersection is still evident in (incorrect) student response patterns in both the control and (to a somewhat lesser extent) the treatment versions of the electric potential graph task. Thus, while the specific context of electric potential difference appears to strengthen the impact of the inclusion of correct reasoning elements on student performance, it was not sufficient to eliminate the domain-general (or context-independent) intersection responses. In fact, the lowest prevalence of this answer was 17%—still a substantial amount.

TABLE IV. Answer distribution data for the isomorphic graph tasks in chaining format (experiment 1B) for those students who answered both screening questions correctly. The task themselves are shown in the Appendix. (Data from magnetic flux graph task are drawn from the same year as the other three tasks). The correct answer choice is in boldface for reference.

	Kinematics	Potential energy	Electric potential	Magnetic flux
N	122	38	76	90
Time A (correct)	**63%**	**58%**	**75%**	**73%**
Time B (intersection)	22%	34%	17%	22%
Time C	0%	3%	1%	0%
Never	15%	5%	7%	4%

Through the use of the screening questions, in combination with the chaining versions of all four isomorphic graph tasks, we were able to ascertain that the predicted process 1 default answer was still present and quite prevalent (17% to 34%) among all four final answer distributions—even among those who answered both screening questions correctly and were given the relevant conceptual information in the chaining task. In other words, a substantial percentage of students who previously demonstrated the mindware needed to obtain the relevant quantities from a graph and who were provided with reasoning elements that might cue them toward a correct model still answered consistent with a model based on the salient distracting feature.

The observed response patterns occurred across all four different contexts and this pattern was also generally true of students who answered the screening target questions correctly. This suggests that these patterns are driven by domain-general reasoning phenomena rather than stemming from either student difficulties with the relevant concepts and analysis strategies or topic-specific misconceptions. The observed results may stem from a process 1 response that is not followed up with a productive analytic intervention; as such, they are a natural consequence of all reasoning pathways beginning with type-1 processing.

4. Analysis of incorrect reasoning chains: Cross-task comparison

Because the element structures of each task were identical, comparison across tasks is possible. To analyze the reasoning chains of those students who selected the

TABLE V. Incorrect reasoning chain categories (experiment 2B). Values shown are percentages of those responses in support of time B. Total number of time B responses is indicated for each task.

	N	2 or 3 of 3 productive	1 of 3 productive (No. with derived heuristic)	Canonical	Other
Kinematics	43	7%	49% (40%)	44%	0%
Potential energy	39	10%	13% (3%)	64%	13%
Electric potential	20	15%	35% (25%)	30%	20%
Magnetic flux	29	17%	31% (7%)	35%	17%

common incorrect answer, we applied the same criteria discussed in Sec. V A 5. The results are shown in Table V.

Across all four tasks, there was a tendency for those students who answered time B on the chaining versions to endorse elements that were productive to the correct line of reasoning: between one-quarter and one-half of these students endorsed at least one element associated with the correct line of reasoning. Interestingly, the prevalence of the "derived heuristic only" category is larger in the kinematics context compared to the other three tasks. In the three other contexts, students tended to either include two or three of the three elements in the analogous velocity triad or use one of the two independent principles alone. We suspect that this is related to the nature of instruction on the different topics. The heuristic of finding the velocity from the slope of a position vs time graph is more common in introductory physics instruction than, for example, finding the induced EMF from the slope of a magnetic flux vs time graph; instead, when teaching flux, the emphasis is typically on the mathematical relationship from Faraday's law (i.e., $\varepsilon = -d\Phi_B/dt$).

In summary, analysis of the incorrect reasoning chains on the isomorphic chaining tasks provided further support for the prediction that students would likely incorporate productive reasoning elements into reasoning chains in support of incorrect answers despite their logical inconsistency from the perspective of an expert.

C. Summary

In experiment 1A, we utilized the kinematics graph task to investigate the working hypothesis that providing improved access to relevant conceptual information would not cause most students to abandon an initial incorrect model. A variety of measures provided evidence for this hypothesis: namely, comparison of answer distributions between treatment and control, the lack of a sizable difference in answer distributions among those judged to have relevant mindware, and an analysis of incorrect reasoning chains. Analysis of results from the isomorphic graph tasks employed in Experiment 1B also supported the proposed mechanisms (from dual-process theories of reasoning) driving the selection and abandonment of mental models. These data also established that these mechanisms are at play in contexts outside of kinematics. The predictions drawn from the working hypothesis about student answering patterns and behavior when responding to reasoning chain construction tasks were thus found to be applicable not just in context of kinematics, but across four different physics contexts.

VI. EXPERIMENT 2A AND 2B: EXAMINING THE IMPACT ON STUDENT ANSWERING PATTERNS OF A STATEMENT THAT REFUTES AN INCORRECT DEFAULT MODEL IN THE CONTEXT OF A FRICTION TASK

Experiments 1A and 1B demonstrated that providing relevant conceptual information to students did not generally lead to changes in student answer distributions on physics graph tasks. This supported the working hypothesis that an incorrect default model would only be abandoned in the presence of information that casts doubt upon this model. In the sections that follow, we discuss experiments 2A and 2B. In experiments 2A and 2B, we provided information that refutes the default model and explored whether a productive engagement of the analytic system leading to a larger change in student answer distributions occurred. Experiment 2A therefore allowed us to answer RQ2 (impact of statement that refutes default model). We also investigated (in experiment 2B) whether this refutational information affected students differently depending upon their previously demonstrated mindware, thereby answering RQ3 (extent to which impact depends on mindware).

A. Experiment 2A: Refuting the default model

In experiment 2A, students were provided with an element that was intended to stimulate a more productive process 2 intervention by promoting cognitive reflection. Process 2 reasoning often represents an analytic intervention triggered by a low feeling of rightness [31] with the initial model. It is primarily concerned with evaluating satisfaction with the initial model. If the feeling of rightness is strong, the analytic process either may not be engaged (as there is no red flag to prompt efforts to scrutinize the default model) or may be engaged only superficially. To induce a more productive analytic intervention, the feeling of rightness needs to be lowered to a point where the default model becomes unsatisfactory such that increased scrutiny becomes necessary. In experiment 2A, we attempted to decrease the feeling of rightness in the context of the chaining format via a relatively modest

TABLE VI. Reasoning elements provided to the students on the chaining version of the two-box friction task [see Fig. 2(b)]. Elements productive to the correct line of reasoning are bolded. The final two elements had a text box in which students could indicate whether the friction force was *greater than*, *less than*, or *equal to* the applied force for each box, and students were given special instructions on how to use these text boxes. The analytic intervention element, which was present only in the treatment condition, is indicated by an asterisk.

$F_{\rm net} = ma$
Both boxes have the same mass
The tension force on box A is equal to the tension force on box B
Both boxes remain at rest
Coefficient of friction for A is smaller than the coefficient of friction for B
Both boxes have the same weight
The normal force on box A is equal to the normal force on box B
Neither box is accelerating
The horizontal forces are balanced
The vertical forces are balanced
The net force on both boxes is zero
The friction force and the applied force are the only horizontal forces acting on the box
The coefficient of static friction is not relevant to this problem*
$\mathbf{F_{frct\,on\,A}}$ **is [insert relationship here]** $\mathbf{F_{app\,on\,A}}$
$\mathbf{F_{frct\,on\,B}}$ **is [insert relationship here]** $\mathbf{F_{app\,on\,B}}$

intervention; in particular, we inserted a single reasoning element into the list that explicitly refuted the incorrect default model.

1. Description of experiment 2A

Experiment 2A utilized the two-box friction task described in the introduction [see Fig. 2(b) and Ref. [4]] cast into the chaining format. In the two-box friction task, students are asked to compare the magnitudes of the friction forces on two identical boxes on different surfaces. The coefficient of friction for each box-surface pair is indicated on the diagram. These coefficients have empirically been found to function as a salient distracting feature for students, resulting in a common incorrect answer based on reasoning from the coefficients alone [4].

The reasoning elements given to the students in this task are shown in Table VI. Note that two of the reasoning elements allow students to define the relationship between $F_{\rm frct}$ and $F_{\rm app}$ on both boxes. While every other element given to the student contains a true statement, these two elements may or may not be true depending on what the student fills in. The treatment group received the chaining version of the friction task with the element “the coefficients of friction are not relevant to this problem” included. In this article, we refer to this element as the *analytic intervention element,* or AIE, because it was designed to stimulate a more productive analytic intervention by reducing the satisfaction with the model that the coefficients of static friction determine the relative magnitudes of the static friction forces. This element was intended to “nudge” students into scrutinizing the incorrect default model and to encourage them to explore alternative models. The control group received a chaining version of the friction task that did not include the AIE.

2. Predictions

In experiment 2A, the chaining format was used for both the control and the treatment groups. Indeed, we had already found in experiments 1A and 1B that the chaining format itself, which includes reasoning elements productive to a correct line of reasoning, is unlikely to change answering patterns substantially in comparison to a more standard multiple choice with explanation format. However, we expected that the inclusion of the analytic intervention element would reduce satisfaction with the default model and would therefore result in a bigger difference between the treatment and control groups. Thus, our prediction for experiment 2A was that there would be a difference in the answering distributions between the treatment and control conditions.

3. Analysis of answer distributions and discussion

Student answer distributions from both versions (control and treatment) are shown in Table VII. The data were collected in two different semesters (both on- and off-sequence) of the introductory calculus-based mechanics course.

While the overall performance in the on-sequence and off-sequence courses differed substantively, in both trials there was a statistically significant, medium-effect-size difference in answer distributions for the treatment condition with respect to the control condition. This suggests that the AIE had an impact on the answer distributions.

The overall performance difference between the on- and off-sequence courses may stem from some combination of differences in instruction (e.g., the on-sequence course implemented *Tutorials in Introductory Physics* [37] with high fidelity, while the off-sequence course did not) and differences in participation rates and participation incentives among the two courses. The absolute performance difference

TABLE VII. Student answer distributions on both versions (control and treatment) of the chaining version of the two-box friction task (experiment 2A). The task itself is shown in Fig. 2(b). The correct answer choice is in boldface for reference.

	On sequence		Off sequence	
	Control	Treatment	Control	Treatment
N:	119	120	64	66
$\mathbf{F_{frct\ on\ A} = F_{frct\ on\ B}}$ **(correct)**	**55%**	**74%**	**27%**	**38%**
$F_{\text{frct on A}} < F_{\text{frct on B}}$ (common incorrect)	35%	23%	70%	50%
$F_{\text{frct on A}} > F_{\text{frct on B}}$	10%	2%	3%	12%
Not enough info	0%	1%	0%	0%
	$p = 0.003,\ V = 0.22$		$p = 0.04,\ V = 0.23$	

between on sequence and off sequence was of less interest to our investigation than the differences in answer distributions between treatment and control. However, it is worth mentioning that residual analysis of the on-sequence data shows elevated counts in the correct answer for the treatment condition (+3.16 for “equal,” −2.03 for “less than,” and −2.42 for “greater than” or ”not enough info”) while residual analysis of the off-sequence data shows less of a difference in correct answering for the treatment condition (1.38 for “equal”, −2.36 for “less than”, and 1.92 for “greater than”/“not enough info”). These differences in the residuals may relate to the interaction of the AIE with the baseline level of understanding between the two groups; this idea is further addressed in experiment 2B.

Table VII demonstrates that the AIE impacted student answering patterns regardless of the baseline level of understanding (indicated by the performance of the control group from each population). Although differences in performance between the on-sequence and off-sequence groups suggest differences between these two populations, the AIE produced a medium effect-size difference in answer distributions in both groups, with elevated counts for the correct answer in the treatment condition. The fact that we observed a difference in answer distributions in both courses provides further evidence for the generalizability of our results.

4. Analysis of reasoning chains

In our investigation, we examined the reasoning chains constructed by those students in both conditions and categorized them according to a set of criteria described below. Ambiguous responses were discussed by two or more of us until agreement was reached. In this section, we describe these categories and discuss their prevalence, which is indicated in Table VIII.

Most students’ correct responses contained chains that clearly indicated correct reasoning (more than 65% of correct responses in all trials). Generally, these responses included an indication of Newton’s 2nd law being used to determine that the horizontal forces are balanced on both boxes. An example is given below:

> *“both boxes have the same mass / and / the normal force on box A is equal to the normal force on box B / so / because /* $F_{\text{net}} = ma$ */ and / both boxes remain at rest / the horizontal forces are balanced / and / the net force on both boxes is zero / because / the friction force and the applied force are the only horizontal forces acting on the box /* $F_{\text{frct on A}}$ *is equal to* $F_{\text{frct on B}}$*”*

Other correct response chains from students were ambiguous; they could easily be seen as indicating correct

TABLE VIII. Comparison of reasoning chain categories in experiment 2A for on-sequence and off-sequence courses. Percentages shown represent percentage of students in the respective column. (In the text, percentages of correct or incorrect responses were reported for ease of discussion).

	On sequence		Off sequence	
Target question	Control	Treatment	Control	Treatment
N	119	120	64	66
Correct w/ correct reasoning	44% (52)	58% (69)	17% (11)	26% (17)
Ambiguous correct reasoning	7% (8)	8% (9)	3% (2)	9% (6)
Other correct	1% (1)	2% (2)	0% (0)	3% (2)
No reasoning given	3% (4)	8% (9)	6% (4)	0% (0)
Canonical incorrect reasoning	20% (24)	14% (17)	31% (20)	29% (19)
Conceptual difficulty incorrect reasoning	6% (7)	4% (5)	22% (14)	14% (9)
Struggle reasoning	2% (2)	3% (3)	6% (4)	5% (3)
Other incorrect	3% (3)	0% (0)	2% (1)	2% (1)
No reasoning given	15% (18)	5% (6)	13% (8)	14% (9)

reasoning but could also possibly be interpreted as rationalization based on the features of the problem that are equal. For example, one student responded as follows:

> *"$F_{\text{frct on A}}$ is equal to $F_{\text{frct on B}}$ / because / both boxes remain at rest / and / the tension force on box A is equal to the tension force on box B"*

Incorrect reasoning chains were also classified into several common categories. The most prevalent of these categories is represented by the following chain:

> *"both boxes have the same mass / but / coefficient of friction for A is smaller than the coefficient of friction for B / so / $F_{\text{frct on A}}$ is less than $F_{\text{frct on B}}$"*

This student responded with a "canonical" incorrect answer—an answer that primarily relies on a direct judgment based on the coefficients of friction or the equation $f = \mu N$ without reference to other physics principles. Around half of the incorrect responses (between 43% and 55%) fell into this category in each semester, in both treatment and control conditions.

Other incorrect responses utilized the coefficient reasoning but included other pieces of relevant information such as the observation that the boxes remained at rest. For example, one student argued:

> *"both boxes have the same weight / and / the normal force on box A is equal to the normal force on box B / but / neither box is accelerating / because / both boxes remain at rest /and / coefficient of friction for A is smaller than the coefficient of friction for B / therefore / $F_{\text{frct on A}}$ is less than $F_{\text{frct on B}}$"*

This response seems to be consistent with an incorrect conception in which friction is greater than the applied force until the applied force is big enough to *overcome* that friction force. Thus, the friction forces can differ from one another but still be larger than the applied force, thereby leading to both boxes remaining at rest. Here, the student did not appear to answer purely based on the coefficients alone but tried to reconcile the coefficient reasoning with other knowledge about forces. The student likely had some form of process 2 engagement, although one that resulted in an erroneous justification possibly serving to rationalize a default answer. We called this category "Conceptual Difficulty Incorrect Reasoning."

Other students gave responses similar to the following:

> *"the normal force on box A is equal to the normal force on box B / and / both boxes have the same weight / but / coefficient of friction for A is smaller than the coefficient of friction for B / so / Custom: "B needs more force to move" / but / Custom: "since neither of them moved" / the horizontal forces are balanced / and / neither box is accelerating / and / the net force on both boxes is zero / therefore / both boxes remain at rest / but / Custom: "since the coefficient of friction for B is greater" / $F_{\text{frct on A}}$ is less than $F_{\text{frct on B}}$"*

This response shows a student who appeared to struggle between a desire to incorporate correct knowledge and a desire to hold fast to a strong default model, similar to the incorrect responses we saw on the isomorphic graph tasks. In virtually every case, such responses made use of the element "the horizontal forces are balanced" along with accompanying information about Newton's 2nd law. These responses, however, were much less prevalent (<10% of incorrect responses in the on-sequence course, and <10% for the off-sequence course) for the two-box friction task in the two semesters in which experiment 2A was implemented than they were for the graph tasks. Hence, we did not attempt to establish and evaluate such responses according to rigorous criteria in order to determine upper and lower bounds on the extent to which this type of struggle was occurring for students; instead, we opted to identify them as we would other reasoning categories via consensus among us.

Overall, the findings from our analysis of the incorrect responses are consistent with dual-process theories of reasoning. In the context of our framework, those who are attracted to the salient distracting feature (the coefficients) likely have a strong feeling of rightness in a model of friction associated with the coefficients, resulting in a low motivation to search for alternate models. The most prevalent reasoning chains leading to an incorrect answer among all students was the canonical category, with no indication of any reflection on the answer beyond a single model built around the coefficients. There is also an interesting interaction with baseline level of understanding as indicated by the greater prevalence of responses in the conceptual difficulty category in the off-sequence responses; this interaction was explored in greater detail in experiment 2B.

B. Experiment 2B: Testing the effect of mindware

In our working hypothesis, we stated that a productive analytic intervention would require both some level of dissatisfaction with the incorrect default model as well as the mindware necessary for a correct model. In experiment 2A, it was demonstrated that an element that confronted student satisfaction with the incorrect default model successfully altered answering patterns on the two-box friction task, with elevated counts for the correct answer in the treatment condition. In experiment 2B, we modified experiment 2A to test the full extent of the working hypothesis with a focus on the need for mindware supporting a correct model. Experiment 2B therefore enabled us to answer RQ3 (extent to which impact depends on relevant mindware).

TABLE IX. Response data for the two-box friction task separated into control (no AIE) and treatment (with AIE) groups while controlling for performance on the screening question (experiment 2B). The task itself is shown in Fig. 2(b). The correct answer choice is in boldface for reference.

	Screening correct (with correct reasoning)		Screening incorrect	
	Control	Treatment	Control	Treatment
N:	40	39	41	46
$\mathbf{F_{frct\,on\,A} = F_{frct\,on\,B}}$ **(correct)**	**60%**	**90%**	**39%**	**41%**
$F_{\mathrm{frct\,on\,A}} < F_{\mathrm{frct\,on\,B}}$ (common incorrect)	40%	8%	56%	54%
$F_{\mathrm{frct\,on\,A}} > F_{\mathrm{frct\,on\,B}}$	0%	2%	3%	5%
Not enough info	0%	0%	2%	0%
	$p = 0.001,\ V = 0.39$		$p = 1,\ V = 0.025$	

1. Description of experiment 2B

To gauge the effect of having the requisite mindware for a correct model, we repeated experiment 2A with a single modification: the screening question originally used before the two-box friction task by Kryjevskaia *et al.* [4] was administered to students in both conditions immediately before they were given the chaining version of the two-box friction task.[3] We thus operationalized student possession of the requisite mindware as answering the screening question correctly with a correct explanation. This allowed us to probe the impact of the analytic intervention element on students who did and did not possess the requisite mindware by controlling for performance on the screening question.

2. Predictions

We expected that a difference in answering patterns would be more likely to occur for those students who possessed the relevant mindware necessary to replace the default model with something more satisfactory. Without such mindware, the default model would likely be ratified by process 2 because of its initial salience and associated feeling of rightness [41–43]. Thus, our prediction for experiment 2B was that any shift caused by the analytic intervention element would primarily manifest itself in the responses of those students who answered the screening question correctly.

3. Analysis of answer distributions and discussion

Results are shown in Table IX. Among students who demonstrated appropriate mindware, there is a significant difference, with a large effect size, in the answer distributions between students in the treatment group and those in the control group ($p = 0.001$, $V = 0.39$). Residual analysis indicates the difference is due to a deviation from expected counts in the common incorrect (-3.36) and the correct ($+3.04$) answer choices. (The residual for the *other* category was 1.02.) Among students who did not demonstrate appropriate mindware, answer distributions on the two-block friction question were virtually indistinguishable between treatment and control groups ($p = 1$, $V = 0.025$). These findings therefore support our original prediction. By and large, students who demonstrated that they possessed the relevant mindware and had access to the AIE answered the target question correctly. Those students who did not demonstrate that they possessed the relevant mindware, on the other hand, did not gain any benefit from the AIE, as predicted by dual-process theories of reasoning.

Consistent with previously published research on this task [4], many students responded correctly on the screening question but went on to answer the target question incorrectly when the AIE was not present. These results suggest that some students who had the requisite mindware available to them may have been prevented from applying that knowledge on the target question because of a strong feeling of rightness associated with an incorrect default model cued by the salient distracting feature. When this feeling of rightness was challenged by the AIE, such students may have been able to engage in cognitive reflection and arrive at a correct answer using the appropriate mindware. However, students who did not have the requisite mindware available to them were unaffected by the AIE because they did not have the mindware necessary to replace the default model with a more satisfactory alternative model.

4. Analysis of reasoning chains

Table X gives a breakdown of reasoning chains for the target question while controlling for performance on the screening question. Each response was categorized based on the nature of the reasoning presented using the categories described in Sec. VI A 4. In general, a similar pattern emerges as was seen in experiment 2A: most correct answers were accompanied by correct reasoning, and about half of the students who chose the common incorrect answer employed reasoning that only references the single

[3] Unlike in experiments 1A and 1B, the screening question was placed immediately before the target question in part to replicate the way the screening-target pair had been administered in free-response format by Kryjevskaia *et al.* [4].

TABLE X. Comparison of reasoning chains in experiment 2B controlling for performance on the screening question shown in Fig. 2(a). Percentages shown represent percentage of students in the respective column.

	Screening correct (with correct reasoning)		Screening incorrect	
Target question	Control	Treatment	Control	Treatment
N	40	39	41	46
Correct w/ correct reasoning	50% (20)	82% (32)	17% (7)	24% (11)
Ambiguous correct reasoning	10% (4)	8% (3)	20% (8)	13% (6)
Other correct	0% (0)	0% (0)	2% (1)	4% (2)
Canonical incorrect reasoning	20% (8)	3% (1)	34% (14)	35% (16)
Conceptual difficulty incorrect reasoning	8% (3)	5% (2)	5% (2)	13% (6)
Struggle reasoning	13% (5)	0% (0)	12% (5)	0% (0)
Other incorrect	0% (0)	3% (1)	10% (4)	11% (5)

model based on the coefficients (*canonical incorrect*). However, when controlling for performance on the screening question, a new pattern emerges. Using a 7 × 2 Fisher's exact test to compare the treatment and control conditions, it is seen that for the screening-correct population, the prevalence of reasoning categories is statistically different ($p = 0.001$).[4] Furthermore, examining the residuals reveals that the prevalence of correct reasoning has higher than expected counts in the treatment condition (residual was $+3.00$), while the canonical and struggle reasoning categories have lower than expected counts (residuals of -2.44 and -2.28, respectively). For the screening incorrect population, a 7 × 2 Fisher's exact test reveals no difference between the two conditions (control and treatment, $p = 0.20$, $V = 0.31$). However, residual analysis shows that the struggle category has lower than expected counts in the treatment condition (residual -2.44).

These results are consistent with a dual-process perspective of the reasoning dynamics. The AIE is expected to refute the common incorrect default model that cues the canonical reasoning. Those students who have the relevant mindware (i.e., the ability to construct a correct reasoning chain if not cued on an incorrect model) and are prohibited from using it by an incorrect default model would be expected to be most helped by the AIE. It is of note, therefore, that based on the residuals, the canonical and struggle categories show the most decrease in prevalence in the screening correct population, and that the conceptual difficulty category did not seem to change in the presence of the AIE. Those students who have an incorrect default model cued by the SDF and also possess correct mindware (whether or not they explicitly struggle in reconciling the two or not) need only to have their feeling of rightness in that model diminish before they would be able to replace the default model with the correct model and assemble the correct reasoning. This might also explain why, in the screening incorrect condition, those who have incorrect or incomplete mindware may have a diminished feeling of rightness but be unable to pivot towards correct reasoning, opting instead to reconcile their incorrect default model with other incorrect conceptual knowledge.

Students in the control condition who used correct reasoning on the screening question and responded to the target question incorrectly with chains that fell into the canonical incorrect category or the struggle incorrect category were likely inhibited from using the requisite mindware due to the cueing of an incorrect default model by process 1. We argue that if these students had access to the AIE in their reasoning elements, they may have engaged in cognitive reflection and overcome the initial feeling of rightness in this incorrect default model, ultimately responding with correct reasoning after a productive process 2 intervention.

C. Summary

In experiments 2A and 2B, we utilized the two-box friction task to investigate the working hypothesis that presenting students with information that refutes an incorrect default model would cause more students to abandon that model. The comparison of the prevalence of common incorrect answers between control and treatment groups shows that the presence of the AIE reduced the number of common incorrect answers. Additionally, experiment 2B tested the second part of the working hypothesis that those students who possessed relevant mindware would be differentially impacted by statements that refute the default model. This prediction also proved to be correct—in experiment 2B, those students who answered the screening question correctly seem to be positively impacted by the AIE to a greater degree than those who answered the screening question incorrectly. This was seen not only in prevalence of correct or incorrect answers, but also in the nature of the reasoning chains presented by students in defense of their answers.

[4] An effect size was unable to be calculated due to the "other correct" category having zero counts in both the comparison conditions. However, removing that category, the effect size would have been $V = 0.42$.

VII. CONCLUSIONS AND NEXT STEPS

The overarching aim of this investigation was to study the extent to which dual-process theories of reasoning could account for reasoning phenomena on qualitative physics questions using a new methodology involving reasoning chain construction tasks. In particular, we drew upon dual-process theories of reasoning to make and test predictions about student behavior on these chaining tasks. From Evans' heuristic-analytic theory, we developed a working hypothesis that students would be unlikely to shift away from an incorrect default model generated by process 1 unless they were provided with information that explicitly challenges satisfaction with that model and possessed the relevant correct conceptual knowledge (mindware). Two sets of experiments built on the chaining task methodology were devised to test this hypothesis. In the first, students were given graph tasks with a known salient distracting feature [the intersection point, see Fig. 1(b)] which had been cast into a chaining format; the reasoning elements in the chaining task version of the graph task functioned to give students increased access to relevant conceptual information, thus testing whether or not this improved access would be sufficient to impact student answering patterns. In the second set of experiments, we gave students access to information (via the analytic intervention element, or AIE) that could challenge a common incorrect default model about static friction in order to determine whether the presence of this information impacts student answering patterns. We also used a screening question to examine the extent to which the impact of the AIE depended on whether or not students possessed mindware supporting a correct model.

The first set of experiments demonstrated that, in the presence of a salient distracting feature, providing increased access to relevant, correct information does not substantially alter student response patterns. Experiment 1A showed this in the context of a kinematics question and illustrated that information that an expert would consider correct and relevant to the correct response was used by many students to justify an incorrect (and therefore inconsistent) answer. In experiment 1B, the results were reproduced in two other content domains.

The second set of experiments demonstrate that a large difference in answering patterns could in fact be realized by providing access to information that could challenge a common incorrect default model cued by a salient distracting feature. In addition, it was also revealed that this effect was limited to students who had previously demonstrated relevant mindware.

The results from all experiments provide support for the use of dual-process theories as a mechanistic framework for making and testing predictions about student responses and behavior—particularly about which models are selected and why some are abandoned.

This work also has some broader implications related to the interplay between conceptual understanding and reasoning skills. Indeed, our research strongly suggests that, as outlined by dual-process theories of reasoning, process 1 serves as the entry point to any reasoning pathway. As a result, both the nature of the default model generated by process 1 and the way in which students subsequently interact with that model (e.g., using it to reason or evaluating its appropriateness) can strongly impact student responses. In our investigation, we found evidence that those students who possess the relevant mindware to answer a problem correctly may not use that mindware due, in part, to a failure to adequately scrutinize an intuitively appealing default model from process 1. Given that these students demonstrated that they possessed relevant mindware that could both be used to refute the default model and to assist with the generation of a new, normative model, our work suggests that either they may not have fully developed the skill to critically reflect on an intuitive model cued by process 1 (i.e., engage in cognitive reflection), or they may not have incorporated the practice of cognitive reflection into reasoning in physics. One could employ the cognitive reflection test developed by Frederick [33] to measure a student's propensity for cognitive reflection and, using that measure, further investigate whether students who were successful in the control condition demonstrated stronger cognitive reflection skills and how cognitive reflection skills interacted with the impact of the provided reasoning elements, but this is beyond the scope of the current investigation. Here, we argue that cognitive reflection productive type 2 processing may be cued, regardless of disposition, by information refuting a default model. Moreover, our results suggest that the AIE, which was designed to promote cognitive reflection, had no impact on students who did not possess the relevant mindware in experiment 2B. Thus, it is quite likely that students need a certain base level of mindware pertaining to a topic before being able to fully and productively employ cognitive reflection and type 2 processing in order to arrive at a normative response.

It is clear from the current investigation (and others reported in the literature) that domain-general reasoning skills affect the process of content-specific reasoning, and that physics instructors should attend to the development of both domain-general reasoning skills and content-specific mindware if improved performance is a goal. More work is needed to characterize with greater resolving power the interplay between both factors in order to provide detailed research-based approaches for supporting reasoning skills and conceptual understanding in a more integrated fashion. Indeed, probing the impact of domain-general cognitive reflection skills on student reasoning and performance in physics has been the focus of recent and ongoing work by our research team (see, for example, Refs. [44,45]). Given the nature of our current findings, however, it is important for instructors to recognize that poor performance on a specific physics task may not be indicative of a lack of relevant conceptual understanding (or mindware) and may, in some cases, be attributed to domain-general reasoning phenomena and students' reasoning skills (including, for example, their cognitive reflection skills). Thus, we recommend that instructors use suites of questions targeting a given topic and be mindful of whether or not a given question contains a likely salient distracting feature. In addition, it may be beneficial for instructors to discuss the dual nature of human reasoning explicitly and to encourage students to scrutinize first-available mental models, particularly those that do not seem to be rooted in physics concepts, and to check them against relevant physics concepts and laws (e.g., Newton's 2nd law).

The successful leveraging of dual-process mechanisms in this work suggests a possible pathway to develop the skills needed to overcome an incorrect default model cued by a salient distracting feature. To be clear, the current work does not demonstrate or investigate long-term changes in the performance of students who were given the AIE, though we recognize the importance of such long-term studies. Instead, it demonstrates the efficacy of an in-the-moment intervention. Giving students access to information that could challenge the default model apparently caused students to scrutinize the default model and to explore and evaluate other relevant physics models during the time period in which they were answering the two-box friction task. If this scaffolded prompting to search for other models could be repeated on many tasks with salient distracting features over a period of one or two semesters, students may begin to internalize a prompt to reflect on intuitive (process 1) answers. While this scaffolding could be provided directly by a line of questioning on a specific tutorial worksheet, it may also be the case that more "hidden" scaffolding (e.g., that provided by the AIE) could be more effective in that, by interacting with the AIE, students are recognizing and modifying their answers without explicitly being prompted to do so. At some point, however, we suspect that students should be explicitly instructed about the impact of salient distracting features on student reasoning and how engaging in cognitive reflection and searching for alternate answers can improve decision-making when these features are present. Based on the present work, we do not believe (and definitely do not claim) that a single intervention would be sufficient to prevent students from making similar mistakes on similar question in the future (for which intervention elements are not provided). One could imagine, however, that it might be productive to have students reflect systematically on their interaction with an AIE after a specific task is completed. We suspect that instruction of this sort may aid students in developing the reflective skills necessary to effectively navigate qualitative physics questions with salient distracting features. More research, of course, is needed to gain insight into the effectiveness of specific pedagogical approaches.

Finally, we argue that this investigation has illustrated the power of reasoning chain construction tasks in exploring the kinds of domain-general reasoning phenomena predicted by dual-process theories of reasoning. Indeed, the mechanisms put forward by these theories can be used to make and test predictions about patterns in student responses, and chaining tasks can readily be manipulated to isolate, to the extent possible, specific mechanisms. We anticipate that the results of studies such as the one reported in this paper, which employ novel methodologies combined with the dual-process framework, can be leveraged to improve the learning and teaching of physics more broadly.

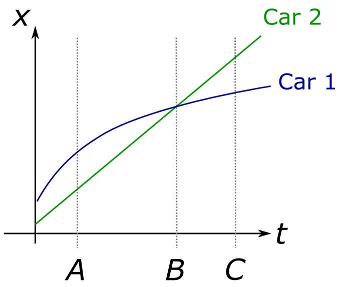
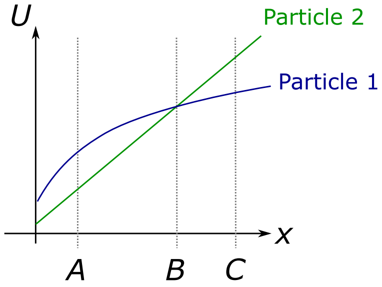
ACKNOWLEDGMENTS

The authors wish to acknowledge valuable discussions with and contributions from Paula R. L. Heron, Andrew Boudreaux, and Andrew F. Heckler. The authors also appreciate substantive feedback about this work from John R. Thompson, Michael C. Wittmann, Natasha M. Speer, past and present members of the Physics Education Research Laboratory at the University of Maine, and two anonymous reviewers. In addition, the lead author would like to acknowledge the University of Maine for support via a Signature and Emerging Areas of Excellence Graduate Fellowship in the area of STEM Education. This material is based upon work supported by the National Science Foundation under Grants No. DUE-1431940, No. DUE-1431541, No. DUE-1431857, No. DUE-1432052, No. DUE-1432765, No. DUE-1821390, No. DUE-1821123, No. DUE-1821400, No. DUE-1821511, No. DUE-1821561, and No. DRL-0962805.

APPENDIX: ISOMORPHIC GRAPH TASKS

The task statements, reasoning elements, and screening questions for the four isomorphic graph task questions are included here for reference.

A. Task statements

Task	Kinematics Graph Task	Potential Energy Graph Task
Figure		
Task Statement	The motions of two cars are described by the position vs. time graphs shown above. When, if ever, are the magnitudes of the velocities (i.e., the speeds) of the cars the same?	The potential energy of system 1, in which only particle 1 can move, is described by the potential energy vs. position graph shown. Likewise, the potential energy of system 2, in which only particle 2 can move, is shown. The two systems don't interact. Where, if anywhere, are the magnitudes of the forces on the particles the same?

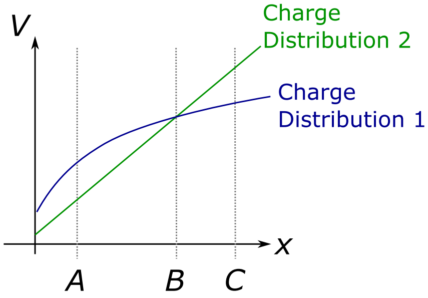
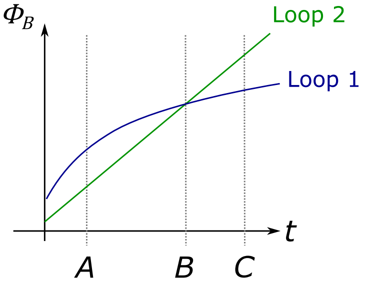
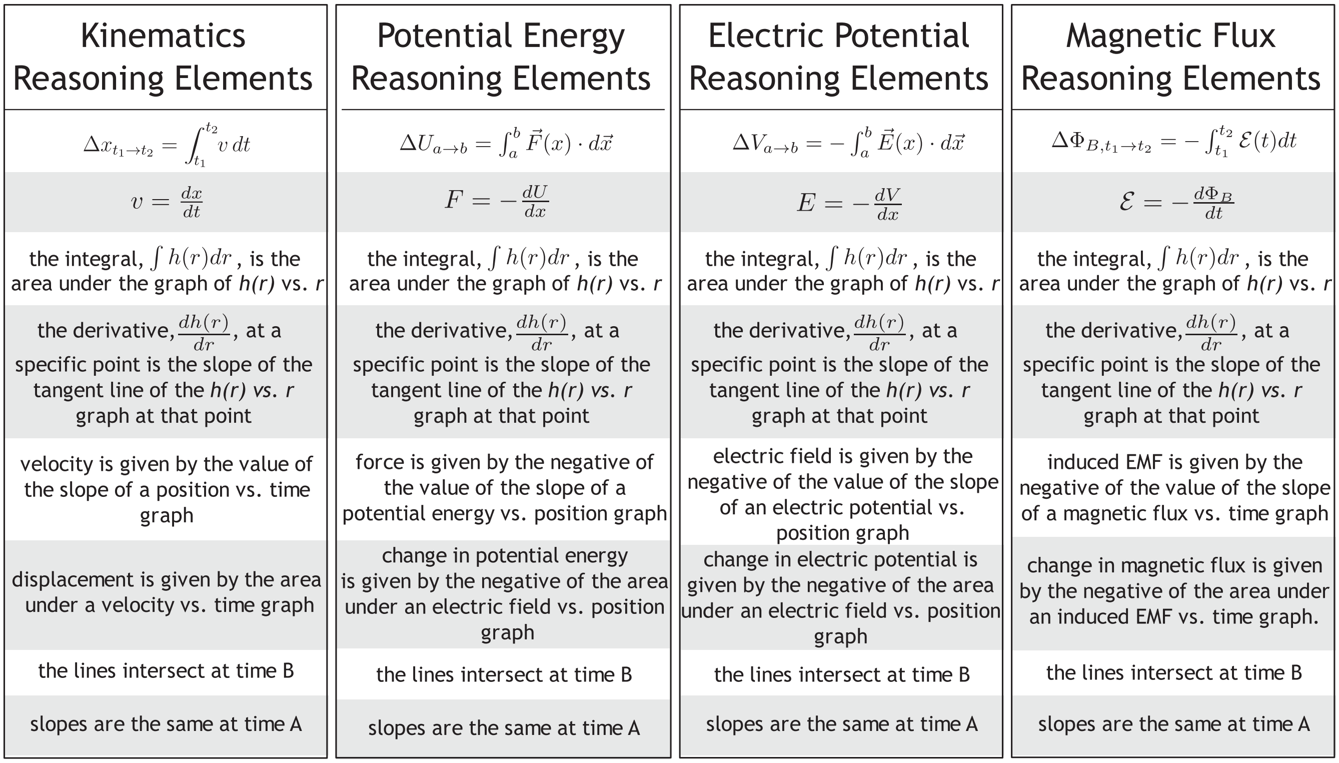
Task	Electric Potential Graph Task	Magnetic Flux Task
Figure		
Task Statement	The electric potentials set up by two charge distributions located far away from each other are described by the electric potential vs. position graphs shown above. Where, if anywhere, are the magnitudes of the electric fields due to the charge distributions the same?	The magnetic fluxes through two different conducting loops in different magnetic fields are described by the magnetic flux vs. time graphs shown above. When, if ever, are the absolute values of the induced EMF's ($\mathcal{E}_1$ and $\mathcal{E}_2$) the same?

FIG. 7. Tasks statements from the four isomorphic graph tasks used in the study.

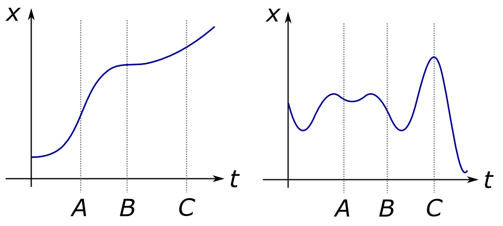
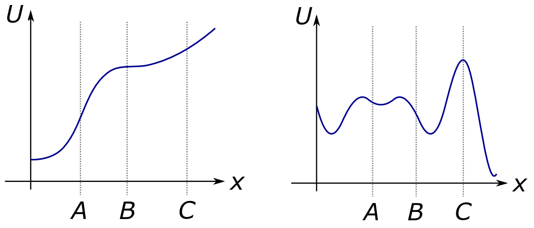
B. Reasoning elements provided



Kinematics Reasoning Elements	Potential Energy Reasoning Elements	Electric Potential Reasoning Elements	Magnetic Flux Reasoning Elements
$\Delta x_{t_1\to t_2} = \int_{t_1}^{t_2} v\,dt$	$\Delta U_{a\to b} = \int_a^b \vec{F}(x)\cdot d\vec{x}$	$\Delta V_{a\to b} = -\int_a^b \vec{E}(x)\cdot d\vec{x}$	$\Delta\Phi_{B,t_1\to t_2} = -\int_{t_1}^{t_2} \mathcal{E}(t)dt$
$v = \frac{dx}{dt}$	$F = -\frac{dU}{dx}$	$E = -\frac{dV}{dx}$	$\mathcal{E} = -\frac{d\Phi_B}{dt}$
the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*	the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*	the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*	the integral, $\int h(r)dr$, is the area under the graph of *h(r)* vs. *r*
the derivative, $\frac{dh(r)}{dr}$, at a specific point is the slope of the tangent line of the *h(r) vs. r* graph at that point	the derivative, $\frac{dh(r)}{dr}$, at a specific point is the slope of the tangent line of the *h(r) vs. r* graph at that point	the derivative, $\frac{dh(r)}{dr}$, at a specific point is the slope of the tangent line of the *h(r) vs. r* graph at that point	the derivative, $\frac{dh(r)}{dr}$, at a specific point is the slope of the tangent line of the *h(r) vs. r* graph at that point
velocity is given by the value of the slope of a position vs. time graph	force is given by the negative of the value of the slope of a potential energy vs. position graph	electric field is given by the negative of the value of the slope of an electric potential vs. position graph	induced EMF is given by the negative of the value of the slope of a magnetic flux vs. time graph
displacement is given by the area under a velocity vs. time graph	change in potential energy is given by the negative of the area under an electric field vs. position graph	change in electric potential is given by the negative of the area under an electric field vs. position graph	change in magnetic flux is given by the negative of the area under an induced EMF vs. time graph.
the lines intersect at time B	the lines intersect at time B	the lines intersect at time B	the lines intersect at time B
slopes are the same at time A	slopes are the same at time A	slopes are the same at time A	slopes are the same at time A

FIG. 8. List of reasoning elements provided for each of the four isomorphic graph tasks.

C. Screening question task statements

Task	Kinematics Screening Questions	Potential Energy Screening Questions
Figures		
Task Statement	The motion of a car is described by the position vs. time graph shown above. At which of the three labeled times is the magnitude of the velocity (i.e., the speed) of the car the greatest?	The potential energy of a system in which only one particle can move is described by the potential energy vs. position graph shown. At which of the three labeled positions is the magnitude of the force on the particle the greatest?

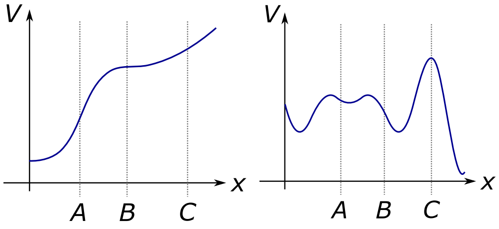
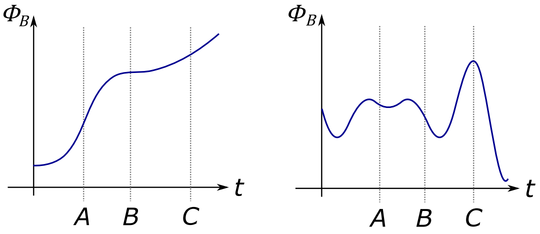
Task	Electric Potential Screening Questions	Magnetic Flux Screening Questions
Figure		
Task Statement	The electric potential set up by a charge distribution is described by the electric potential vs. position graph shown above. At which of the three labeled positions is the magnitude of the electric field due to the charge distribution the greatest?	The magnetic flux through a conducting loop is described by the magnetic flux vs. time graph shown above. At which of the three labeled positions is the absolute value of the induced EMF the greatest?

FIG. 9. Screening question task statements for the four isomorphic graph tasks used in the study.

[1] N. D. Finkelstein and S. J. Pollock, Replicating and understanding successful innovations: Implementing tutorials in introductory physics, Phys. Rev. ST Phys. Educ. Res. **1**, 010101 (2005).

[2] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, Proc. Natl. Acad. Sci. USA **111**, 5 (2014).

[3] 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).

[4] M. Kryjevskaia, M. R. Stetzer, and T. K. Lê, Failure to engage: Examining the impact of metacognitive interventions on persistent intuitive reasoning approaches, in *Proceedings of the 2014 Physics Education Research Conference, Minneapolis, MN* (AIP, New York, 2015).

[5] A. F. Heckler, The role of automatic, bottom-up processes: In the ubiquitous patterns of incorrect answers to science questions, Psychol. Learn. Motiv. **55**, 227 (2011).

[6] A. F. Heckler and T. M. Scaife, Patterns of response times and response choices to science questions: The influence of relative processing time, Cogn. Sci. **39**, 9 (2014).

[7] 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).

[8] C. R. Gette, M. Kryjevskaia, M. R. Stetzer, and P. R. 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).

[9] 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).

[10] D. Kahneman, *Thinking, Fast and Slow* (Mac Millan USA, New York, 2013).

[11] M. McCloskey, *Mental Models*, edited by D. Gentner and A. Stevens (Erlbaum, Hillsdale, NJ, 1983).

[12] G. J. Posner, K. A. Strike, P. W. Hewson, and W. A. Gertzog, Accommodation of a scientific conception: Toward a theory of conceptual change, Sci. Educ. **66**, 4 (1982).

[13] A. A. DiSessa, Toward an epistemology of physics, Cognit. Instr. **10**, 105 (1993).

[14] D. Hammer, A. Elby, R. Scherr, and E. F. Redish, *Transfer of learning: Research and perspectives*, edited by J. Mestre (Information Age Publishing Inc., Charlotte, NC, 2005).

[15] D. Hammer, Student resources for learning introductory physics, Am. J. Phys. **68**, S52 (2000).

[16] P. R. L. Heron, Empirical investigations of learning and teaching, part I: Examining and interpreting student thinking, in *Proceedings of the International School of Physics, Enrico Fermi, No. 156* (Italian Physical Society, Bologna, 2004).

[17] L. C. McDermott, Millikan Lecture 1990: What we teach and what is learned—Closing the gap, Am. J. Phys. **59**, 301 (1991).

[18] L. C. McDermott, Oersted Medal Lecture 2001: Physics Education Research—The Key to Student Learning, Am. J. Phys. **69**, 1127 (2001).

[19] A. Elby, What students' learning of representations tells us about constructivism, J. Math. Behav. **19**, 481 (2000).

[20] P. R. L. Heron, Testing alternative explanations for common responses to conceptual questions: An example in the context of center of mass, Phys. Rev. Phys. Educ. Res. **13**, 010131 (2017).

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

[22] J. S. B. T. Evans and K. E. Stanovich, Dual-process theories of higher cognition, Perspectives Psychol. Sci. **8**, 5 (2013).

[23] T. K. Lê, Using Contrasting Cases to Build Metacognitive Knowledge About the Impact of Salient Distracting Features in Physics Problems, Dissertation, University of Maine (2017).

[24] S. Mamede, T. Splinter, T. van Gog, R. Rikers, and H. Schmidt, Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes, BMJ Quality Safety **21**, 4 (2012).

[25] L. C. McDermott, M. L. Rosenquist, and E. H. Zee, Student difficulties in connecting graphs and physics: Examples from kinematics, Am. J. Phys. **55**, 503 (1987).

[26] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).

[27] 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).

[28] J. C. Speirs and W. N. Ferm Jr., M. R. Stetzer, and B. A. Lindsey, Probing student ability to construct reasoning chains: A new methodology, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA* (AIP, New York, 2016).

[29] A. Elby, Helping physics students learn how to learn, Am. J. Phys. **69**, S54 (2001).

[30] E. F. Redish, A theoretical framework for physics education research: Modeling student thinking, in *Proceedings of the International School of Physics, Enrico Fermi, No. 156* (2004).

[31] V. A. Thompson, Dual-process theories: A metacognitive perspective, in *In Two Minds: Dual Processes and Beyond* (Oxford University Press, New York, 2009).

[32] S. Tishman, E. Jay, and D. N. Perkins, Teaching thinking dispositions: From transmission to enculturation, Theory Into Practice **32**, 147 (1993).

[33] S. Frederick, Cognitive Reflection and Decision Making, J. Econ. Perspect. **19**, 25 (2005).

[34] R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises, Rev. Gen. Psychol. **2**, 175 (1998).

[35] D. N. Perkins, *Outsmarting IQ: The Emerging Science of Learnable Intelligence* (Free Press, New York, NY, 1995); see also K. E. Stanovich, *What Intelligence Tests Miss* (Yale University Press, New Haven, CT, 2010).

[36] Data for this paper were collected using Qualtrics software, Copyright © 2021 Qualtrics. Qualtrics and all other Qualtrics product or service names are registered trademarks or trademarks of Qualtrics, Provo, UT, USA, https://www.qualtrics.com.

[37] L. C. McDermott and P. S. Shaffer, *Tutorials in Introductory Physics* (Pearson College Div, Upper Saddle River, NJ, 2001).

[38] National Research Council, *How People Learn: Brain, Mind, Experience, and School* (National Academies Press, Washington, DC, 1999).

[39] R. D. Knight, *Five Easy Lessons: Strategies for Successful Physics Teaching* (Pearson, Upper Saddle River, NJ, 2002).

[40] A. F. Heckler and E. C. Sayre, What happens between pre- and post-tests: Multiple measurements of student understanding during an introductory physics course, Am. J. Phys. **78**, 768 (2010).

[41] J. G. Johnson and M. Raab, Take The First: Option-generation and resulting choices, Org. Behav. Human Decision Processes **91**, 7 (2003).

[42] A. Tversky and D. Kahneman, Availability: A heuristic for judging frequency and probability, Cogn. Psychol. **5**, 9 (1973).

[43] R. Hertwig, S. M. Herzog, L. J. Schooler, and T. Reimer, Fluency heuristic: A model of how the mind exploits a by-product of information retrieval, J. Exper. Psychol. Learn. Memory Cog. **34**, 1191 (2008).

[44] 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).

[45] 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).