Designing research-based instructional materials that leverage dual-process theories of reasoning: Insights from testing one specific, theory-driven intervention

Mila Kryjevskaia*
Department of Physics, North Dakota State University, Fargo, North Dakota, 58102, USA

MacKenzie R. Stetzer
Department of Physics and Astronomy & Maine Center for Research in STEM Education, University of Maine, Orono, Maine 04469, USA

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

Alistair McInerny
Department of Physics, North Dakota State University, Fargo, North Dakota, 58102, USA

Paula R. L. Heron
Department of Physics, University of Washington, Seattle, Washington 98195, USA

Andrew Boudreaux
Department of Physics, Western Washington University, Bellingham, Washington 98225, USA

(Received 6 July 2019; accepted 24 February 2020; published 4 December 2020)

[This paper is part of the Focused Collection on Curriculum Development: Theory into Design.] Research in physics education has contributed substantively to improvements in the learning and teaching of university physics by informing the development of research-based instructional materials for physics courses. Reports on the design of these materials have tended to focus on overall improvements in student performance, while the role of theory in informing the development, refinement, and assessment of the materials is often not clearly articulated. In this article, we illustrate how dual-process theories of reasoning and decision making have guided the ongoing development, testing, and analysis of an instructional intervention, implemented at three different institutions, designed to build consistency in student reasoning about the application of Newton's 2nd law to objects at rest. By employing constructs from cognitive science associated with dual-process theories of reasoning (such as mindware and cognitive reflection), we were able not only to examine the overall improvement in student performance but also to investigate the impact of the intervention on two aspects of productive reasoning—mindware and cognitive reflection. Our analysis showed that the intervention strengthened students' mindware such that students were able to apply it as a criterion while checking the validity of their intuitive responses. Moreover, logistic regression revealed that the success of our intervention was mediated by the students' cognitive reflection skills. Indeed, for students with comparable mindware, those who demonstrated a weaker tendency toward cognitive reflection were less likely to initiate conflict detection and therefore never had the opportunity to utilize their mindware. We believe that this kind of integrated, theory-driven approach to intervention design and testing represents an important first step in efforts to both account for and leverage domain-general reasoning phenomena in the learning and teaching of physics.

DOI: 10.1103/PhysRevPhysEducRes.16.020140

I. INTRODUCTION

Over the past several decades, research in physics education has contributed to improvements in the learning and teaching of physics by serving as a guide for the development of instructional materials for undergraduate physics courses. Many such research-based curricula have been successfully designed, tested, refined, and adopted nationally and internationally [1–8]. Findings from ongoing

*mila.kryjevskaia@ndsu.edu

research on the effectiveness of these materials continue to reveal impressive learning gains [9–11], even at sites other than the developers' institutions [12–16]. While the details of the development process for these materials are often discussed in the literature, the focus is generally on improvements in overall student performance. In many articles, the theoretical basis for the development is not clearly articulated. As the goals and approaches of research-based curriculum development become increasingly diverse and nuanced, there is a need for detailed descriptions of how decisions about specific instructional interventions are made, as well as of how the developers' underlying theoretical commitments inform the research and assessment efforts that guide the development of their materials.

In this article, we illustrate the ways in which one particular collection of theoretical frameworks, dual-process theories of reasoning and decision making, have informed the ongoing development, testing, and analysis of a single instructional intervention. The intervention focuses on Newton's 2nd law and has been implemented in introductory calculus-based physics courses at three different institutions serving a broad spectrum of students. We foreground the tight connections among theory, intervention design, and assessment. Given the multiple, interrelated learning goals of the intervention, we believe that our process and findings will be of interest to others involved in the design of theory-driven instructional materials as well.

Our curriculum development work has been in direct response to a phenomenon observed by many physics instructors: students often reason inconsistently on questions targeting the same physics concept in somewhat different contexts. Such reasoning inconsistencies have been documented in the literature, and may stem from a variety of factors [17–22]. While some of the observed inconsistencies may arise from a lack of a conceptual understanding, we have been deeply interested in those inconsistencies exhibited by students who otherwise appear to demonstrate a conceptual understanding of a particular topic.

In our prior research, we developed and administered pairs of questions with different surface features but the same underlying conceptual structure. Physics experts quickly recognize that both questions can be answered with the same reasoning approach. In practice, the first question in the sequence requires a relatively straightforward application of one or more key physics principles, whereas the second question is presented in a slightly different context and has been empirically shown to elicit strongly held, intuitively appealing, incorrect responses. While students tend to reason correctly on the first question, many seem to abandon this correct line of reasoning on the second question. Students instead seem to either draw upon intuitive ideas or apply formal knowledge incorrectly in an attempt to justify an intuitively appealing answer. Thus, for many students, domain-general reasoning phenomena appear to inhibit the correct use of the relevant (and already demonstrated) knowledge and skills on that second question.

Researchers in physics education have increasingly turned to findings and models from cognitive science in order to gain insight into such reasoning phenomena [17,18]. In particular, dual-process theories of reasoning and decision making have been used by researchers to explain, in a mechanistic fashion, reasoning inconsistencies similar to those described above. Such theories model human cognition as two interacting, but largely distinct processes: a fast, automatic, subconscious process (process 1, sometimes referred to as the *heuristic process*); and a slow, analytical, effortful process (process 2, sometimes referred to as the *analytic process*) [23,24]. Leveraging these theories, researchers have recognized that productive reasoning relies upon both relevant knowledge of the rules and concepts needed to reason through a given problem (sometimes referred to as *mindware*) and the ability to mediate intuitive thinking by reasoning more analytically (sometimes referred to as *cognitive reflection*) [25,26].

In this paper, we describe the development of a three-stage instructional intervention guided by dual-process theories of reasoning (DPTOR) and the associated constructs of mindware and cognitive reflection. We articulate the role of theory in guiding the design of the intervention as well as the collection, analysis, and interpretation of assessment data before, during, and after instruction. We describe the theory-driven approach we have taken in three distinct phases of the project. In phase 1, we disentangled, to the extent possible, the contributions of mindware and cognitive reflection skills to productive reasoning on one particular physics question. In phase 2, we designed an intervention sequence that supports both the development of relevant mindware and the engagement of cognitive reflection skills. In phase 3, we used assessment data to not only examine the overall improvement in student performance but also to pinpoint the impact of the intervention on *each* aspect of productive reasoning (i.e., mindware and cognitive reflection).

The structure of the paper broadly follows the three overarching phases of our project. Section II of this manuscript provides motivation for the current work and an overview of dual-process theories of reasoning and decision-making, mindware, and cognitive reflection. Sections III–V describe the methodology, data collection, and analysis associated with phases 1, 2, and 3, respectively, at one institution. In Sec. VI, we explore the reproducibility of our results by testing the intervention with two additional populations of students, thereby helping us to articulate somewhat broader implications for curriculum development in Sec. VII. Finally, Sec. VIII summarizes our conclusions.

II. MOTIVATION AND THEORETICAL UNDERPINNINGS

The reasoning inconsistencies highlighted in the introduction are likely familiar to instructors at all levels and

constitute an important phenomenon to address. In this section, we provide an overview of relevant theoretical frameworks and constructs from cognitive science that we employed in our investigation of the development of an instructional intervention aimed at helping students build coherence in their reasoning.

A. Inconsistencies in student reasoning

To investigate inconsistencies in student reasoning, we have developed and administered several question pairs, including the one shown in Fig. 1. In question 1, the *block question*, a thin, massless rod is glued to a heavy block at rest on a table. Students are told that the weight of the block is 50 N and that the table is exerting a 30 N force on the block. Students are asked to determine the direction of the force the rod exerts on the block, and to explain their reasoning. By applying Newton's 2nd law, $F_{\text{net}} = 0$ for the block, students can conclude that the rod exerts an upward force of 20 N. Students typically do well on this question with the majority of students giving a correct answer with correct reasoning. (A more detailed discussion of results is found in Sec. III.)

We regard success on the block question as an indicator that a student knows how to apply Newton's 2nd law for an object in the at-rest condition, and thus refer to it as a *screening question* [18]. Question 2, the *magnet question*, requires that students apply the same underlying reasoning in a situation with more complex surface features that tend to elicit intuitively appealing, incorrect responses. (We thus refer to question 2 as a *target question*.) In question 2, a magnet weighing 10 N remains at rest on the side of a refrigerator while a hand exerts an upward force of 6 N. Students are asked to determine the direction of the friction force exerted on the magnet. A physics expert will recognize the magnet question as analogous to the block question; because the magnet remains at rest, the net force on it must be zero, and thus the friction force is upward since the vertical component of the net force is necessarily zero. Student performance on the magnet question, however, is typically much lower.

Of the students who answer the block question correctly, only about one-third answer the target question correctly. Many students who correctly apply relevant concepts on the block question do not seem to apply the same reasoning on the target question.

Several interpretations for the inconsistencies in student responses on the two questions may emerge. The target question, by design, does not look similar to the screening question and involves vertical forces that differ in nature (i.e., frictional force between the magnet and the fridge vs force by the rod). Moreover, the target question contains distracting features (i.e., magnetic and normal forces between the magnet and the surface of the fridge) that are irrelevant to the analysis of the forces acting in the vertical direction. As such, one may argue that student performance on the target question will necessarily be worse than that on the screening. While we agree that worse performance on the target question may not be surprising, the overarching goal of our research is to pinpoint, understand, and address mechanisms contributing to student reasoning difficulties in the presence of correct formal knowledge from instruction. To explain the inconsistency, one may argue that students in the introductory mechanics course do not yet understand the nature of magnetic forces typically covered in the second semester of the introductory sequence. As a result, they would not be able to apply Newton's 2nd law in this situation correctly. Careful analysis of student written responses to the target question, however, reveals that most students who answered the question incorrectly argued that the friction force is

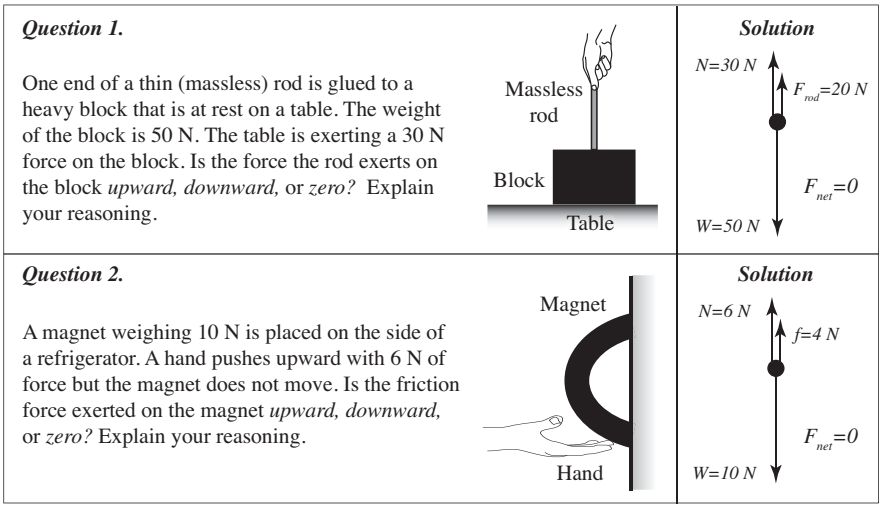


FIG. 1. A pair of questions on forces and Newton's laws. The free-body diagram for question 2 illustrates vertical forces only; horizontal forces acting on the magnet in question 2 are not relevant to the case of static friction and therefore are not included.

downward because it must "oppose" the upward force exerted by the hand. This type of response typically does not make explicit reference to Newton's 2nd law, the gravitational force, or the (irrelevant) magnetic force. This empirical finding is consistent with dual-process theories of reasoning (discussed in detail below), which suggest that the "friction-opposes-hand" argument readily comes to mind due to its ubiquity in many contexts considered in introductory mechanics courses, is intuitively appealing, and may be interfering with students' ability to access and apply physics concepts used previously on the block question [27]. If the presence of the magnetic force is a concern for some students, this concern is not evident from their written responses. It appears that, instead, the students embrace the "friction-opposes-hand" response as highly plausible and do not find the need to analyze the situation any further by applying Newton's 2nd law. An in-depth interpretation of this pattern of reasoning inconsistencies through the lens of DPTOR is included in Sec. III B 2.

A variety of similar reasoning inconsistencies have been reported by us and other investigators [17,18,28]. These results span a variety of conceptual domains in physics and occur both before and after research-based instruction. This body of work suggests that such reasoning inconsistencies are not due solely to a lack of knowledge and are not exclusively associated with a particular type of instruction, but instead are likely related to more fundamental aspects of human reasoning. We take the view that at least some of the mechanisms underlying this phenomenon are domain general. We have therefore begun to draw on certain theoretical frameworks from cognitive science, such as dual-process theories of reasoning, that provide a mechanistic account of reasoning and decision making. We have found that DPTOR can provide insight into observed inconsistencies in student reasoning in physics [18], and have the potential to guide the development of approaches to instruction that might increase coherence in student reasoning [29,30].

B. Intuition and reasoning

Before describing DPTOR in detail, we clarify our use of the terms "intuition" and "intuitive response." Physics instructors often express the desire to help students develop physical intuition. Unfortunately, however, intuition as a construct is somewhat fuzzy. To the best of our knowledge, operational definitions of intuition or intuitive reasoning have not been articulated in the PER literature. As Singh states, "Physical intuition is elusive—it is difficult to define, cherished by those who possess it, and difficult to convey to others. Physical intuition is at the same time an essential component of expertise in physics" [31]. We concur with Singh, but also recognize that experts and novices can differ greatly in the accuracy of outcomes when making decisions based on intuition. We have thus sought a view of intuition capable of spanning the expert-novice continuum. We adopt Kahneman's parsimonious, pragmatic definition of intuition as "nothing more and nothing less than recognition." [23,32] Working from this view, we regard intuition simply as the first-available mental model that is constructed in the mind of a learner when they are confronted by a novel challenge. As we discuss below, forming an initial mental model happens quickly, and generally below the level of conscious awareness. The formation of such a model occurs in response to contextual cues and is based on the learner's repertoire of prior experiences. An expert and a novice physics learner, when presented with the same set of cues in a problem such as the magnet question, may well form different initial mental models due to the vast differences in their relevant prior experience.

We note that "intuitive knowledge" has often been used by physics instructors and physics education researchers to refer to ideas that originate from outside the physics curriculum. In this way, intuitive knowledge has often been framed as being in opposition to formal knowledge (i.e., to the ideas presented in the physics classroom) [33–35]. This view of intuitive knowledge conflicts with the view of intuition as the first-available mental model, in that experts, through repeated practice over many exposures, generally default to the relevant formal physics knowledge as their first-available response. For an expert, then, "intuitive knowledge" and "formal knowledge" are one and the same, whereas for a novice, they often are not. In our research with question pairs, we have documented cases in which the "intuitive response" given by students (i.e., novices) appears to originate from formal knowledge covered in physics class [28,33]. We thus argue that intuition is not limited to everyday knowledge and experiences, but can also stem from formal knowledge. A physics expert, with her extensive repertoire of prior experiences (including formal knowledge), is likely to recognize relevant diagnostic cues in a novel physics problem, and to then form a productive mental model that is well aligned with a normative response. In contrast, a novice, with more limited experience, may key on spurious cues, and form a less productive mental model.

C. Dual-process theories of reasoning and decision making

Since at least the 1960s, cognitive scientists have worked toward domain-general models for human reasoning and decision making. This work has led to the development of a family of frameworks, dual-process theories of reasoning, that successfully accounts for results from a variety of cognitive performance tasks. Here we describe dual-process theories, and discuss how researchers in physics education have used them to account for student reasoning in physics.

Theories in the dual-process family model reasoning with the same basic cognitive architecture, which involves two distinct processes for thinking: one is fast and automatic (process 1), the other is slow and effortful (process 2) [23,24]. Process 1 is always active and operates

beneath the level of conscious awareness. When confronted with a novel situation, process 1 quickly creates a mental model to account for "what is going on." The model is formed on the basis of prior experiences, which are activated in associative memory in response to the features of the situation (i.e., the contextual cues) identified as relevant. Process 1 is the dominant cognitive mode, allowing us to safely cross the road, orient to an exit sign when leaving a building, or answer the question "What is $3 + 2$?" It is important to note that process 1 *cannot* be turned off, and thus the first-available mental model is the entry point in any reasoning chain.

Process 2, on the other hand, is logical and rule based. It occupies our conscious awareness, is capable of computationally expensive serial processing and "cognitive simulation," and is accompanied by a sense of cognitive effort. Process 2 is thus generally experienced as aversive and is engaged only sparingly. Process 2 allows us to solve a jigsaw puzzle, parse a dense text, or answer the question "What is 213×147?"

Most dual-process theories share these same key features [23]. We draw on the heuristic-analytic theory of Evans [24], which refers to processes 1 and 2 as the heuristic and analytic processes, respectively. Figure 2, adapted from Evans [24], illustrates basic cognitive operation in the heuristic-analytic theory. To understand the interplay between these two processes, we apply the heuristic-analytic theory to a hypothetical case of a student encountering a novel physics question. As the student reads the question, the heuristic process generates an initial mental model. According to Evans' theory, only one mental model is considered at a time (*singularity* principle), and that model is selected based on its perceived relevance to the current task (*relevance* principle). The student's heuristic process may focus on contextual cues irrelevant to (or at least not diagnostic of) the normative physics response. These cues, referred to as salient distracting features [17,34,35], may then strongly influence the mental model that is generated. On the magnet question, for example, the student might cue on "a push applied by a hand," which may activate prior life experiences involving efforts to push a stationary object across a carpeted floor. This in turn may lead to an initial mental model along the lines of "friction resists an applied push."

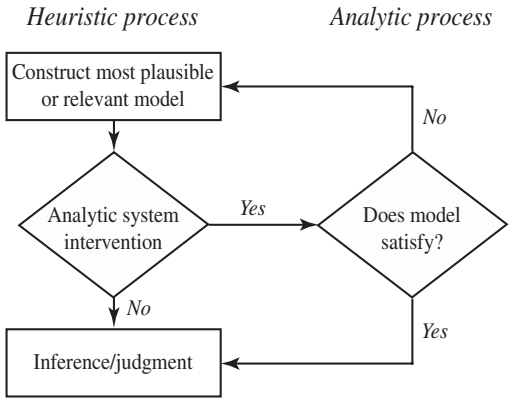


FIG. 2. A diagram illustrating interactions between the heuristic and analytical processes [24].

Once the heuristic process has generated a mental model, the analytic process may or may not intervene. Thompson argues that the default model is accompanied by a value judgment about its plausibility, known as the *feeling of rightness* [36]. If the feeling of rightness is strong and the student is thus confident in his or her approach, an intervention of the analytic process is unlikely; therefore, the analytic process will be bypassed entirely and the student will simply give a response aligned with his or her default mental model [37,38]. In the magnet question example, the student may be familiar with Newton's 2nd law, and the gravitational force, but may not have learned these ideas to a sufficient depth for a "red flag" to be raised (i.e., conflict detection) in association with the answer "friction is downward," generated by the initial model "friction resists an applied push." This direct path from first-available model to response, represented by the pathway on the left side of Fig. 2, is often described as *cognitive miserliness*, because computationally expensive processing is absent [39,40]. Indeed, the reliance on one's "gut feeling" is overwhelmingly prevalent in everyday activities, where it is highly efficient and fairly accurate in guiding judgments.

Many questions posed in physics courses ask students to explain their reasoning. This typically necessitates at least a minimal level of intervention of the analytic process in order to articulate some kind of explanation. Indeed, the job of the analytic process is to ascertain whether or not the default mental model is satisfactory for the task at hand (satisficing principle) [24]. However, even if the analytic process is engaged, it may not necessarily result in a careful examination of the default mental model. In other words, a student might not detect a conflict between this model and formal knowledge, which would potentially lead to an override of the default model in favor of a normative response. Researchers have identified a few mechanisms that impair these processes of conflict detection and subsequent override. First, if a student has a strong feeling of rightness for a particular model, the engagement of the analytic process is likely to be superficial; thus, no conflict detection would be likely. Second, the analytic process is subject to reasoning biases of its own. Reasoners tend to be poor at searching for alternative mental models or generating counterarguments, and thus the analytic process may be driven by confirmation bias [41], thereby shifting its role from conflict detection between the default model and formal knowledge to rationalization of the default model. Third, even if a reasoner attempts to examine whether or not the default model is satisfactory, criteria that a correct response must satisfy may not always be

apparent, especially to a novice learner (even if these criteria are obvious to an expert). As such, an incorrect first-available mental model may still be the final response even after the engagement of the analytic process. This pathway is represented in the lower-right portion of Fig. 2. If, however, the analytic process is engaged, and a conflict with the first-available mental model is detected, the analytic process will hand the task off to the heuristic process for another mental model, as shown in the top-right portion of Fig. 2. Once this new model is generated, the reasoning cycle repeats, which could lead to a sustained override of the first-available mental model in the presence of the adequate relevant mindware.

As illustrated above, the nature of that very first available mental model (which is heavily informed by the student's background knowledge and intuition as well as contextual cues) plays a critical role in the student's overall reasoning process, possibly even precluding the use of relevant knowledge and skills that the student may possess. In order to better characterize the nature of human reasoning and how it may be impacting students as they answer physics questions, we draw upon some other closely related constructs from cognitive science in the section below.

D. Ingredients for productive thinking: Mindware and cognitive reflection

1. Mindware

It is indisputable that relevant knowledge, in the form of rules and procedures, is critical for successful reasoning. In the cognitive science literature, the term mindware, in an analogy to computer software, is used to refer to a collection of "rules, knowledge, procedures, and strategies that a person can retrieve from memory in order to aid decision making and problem solving." [25] Mindware contributes to the formation of the first-available mental model, conflict detection, and the generation of productive alternative mental models (i.e., sustained override). Rather than being simply present or absent, mindware may be instantiated to a greater or lesser depth. Deeper instantiation can support (i) immediate cueing of a first-available mental model consistent with the normative response or (ii) increased conflict detection as well as increased capacity for generating an improved model.

A physics expert will have many modules of physics-related mindware, such as Newton's 2nd law, ingrained much more deeply than a student just learning Newton's laws. A physics expert may know Newton's laws so deeply that the cueing and activation of the heuristic process occurs in accordance with this normative physics knowledge. That is, for the expert, the intuitive, first-available response to the magnet question is likely to be the normative response. However, for students with less robust mindware, if the first-available response is flawed, then conflict detection and override are also likely to be unproductive because threats to the validity of this first available intuitive mental model are less likely to be recognized.

For the magnet question presented in this study, we identified the following pieces of knowledge as constituting the mindware necessary for a successful solution: (i) $F_{\text{net}} = ma$, where F_{net} is the vector sum of all forces, and (ii) for an object at rest (in which case $a = 0$), F_{net} must be zero and thus the vertical component of F_{net} must also be zero. It is, however, challenging to design an instrument (or a task) that determines the degree to which a student possesses this mindware. Several approaches could be considered. For example, a student could be asked to articulate Newton's 2nd law and the conditions necessary for an object to remain at rest. This approach, however, has multiple limitations. First, a student could possess declarative knowledge of Newton's 2nd law and may even memorize the condition for an object to remain at rest. Yet, the student may not be able to recognize the applicability of this knowledge even to a basic scenario that involves more than two forces; in addition, he or she may not be able to identify forces acting on the object or to execute vector addition. For these reasons, we argued that perhaps asking students to consider a basic scenario that requires the application of the identified knowledge is a more informative measure of student mindware. Those students who correctly analyze the scenario are regarded as having at least the minimum level of the necessary mindware. The block question discussed above seems to fit these criteria; it presents a basic situation that involves more than two forces and requires the application of the same set of steps as the magnet question in order to arrive at an answer. Data suggest that only 1 student out of 33 students (∼3%) who answered the block question *incorrectly* provided a correct response with correct reasoning to the magnet question, while ∼70% of the students who answered the block question correctly gave incorrect responses to the magnet question (41 out of 58 students). These results suggest that it is highly unlikely that students will arrive at a correct answer to the magnet question if they failed to do so on the block question. As such, it is reasonable to treat the block question as an adequate instrument for measuring whether or not a student possesses at least the minimum level of the mindware necessary to answer the magnet question correctly; namely, the student recognizes the applicability of Newton's 2nd law to a basic situation involving an object at rest and can successfully apply the two pieces of knowledge articulated above.

2. Cognitive reflection

In addition to mindware, another factor vital for productive reasoning is a general tendency to critically evaluate mental models [26,42]. Differences in relevant mindware notwithstanding, some reasoners are more likely to scrutinize initial models put forward by the heuristic process, and less likely to act solely as cognitive misers.

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half the lake?

FIG. 3. Three-item cognitive reflection test. (The correct answers are 5 cents, 5 minutes, and 47 days, respectively.)

The ability to mediate the heuristic process by reasoning more analytically is typically referred to as cognitive reflection. The reasoner with a stronger disposition toward cognitive reflection will be more likely to engage in conflict detection and subsequent override, which in turn increases the likelihood of reaching a correct answer.

The cognitive reflection test (CRT), developed by Frederick and validated and widely employed in psychology research, has been used to gauge this disposition toward cognitive reflection. [26] Toplak, West, and Stanovich argue that the CRT is "a particularly potent measure of the tendency toward miserly processing," with a lower score indicating increased cognitive miserliness [40]. In the literature, there is a consensus that the CRT is a measure of a reasoner's disposition toward "actively open-minded thinking," (i.e., the tendency to actively search for alternative answers) [26,41,43–45].

The original CRT, shown in Fig. 3, consists of three questions; a revised version contains seven questions [46]. Each question has an intuitively appealing (but incorrect) answer that, for most reasoners, comes to mind immediately and often without conscious effort. To answer correctly, this intuitively appealing response must be tested by the analytic process, deemed unsatisfactory (i.e., a conflict must be detected), and then replaced with a correct answer (i.e., an override process must be successfully competed). For the questions on the CRT, this intervention pathway, if it occurs, typically does so upon only brief reflection, rather than requiring extended analytical processing. This is likely because the CRT items require mindware, such as basic arithmetic, that is relatively strong for most adults (particularly for students enrolled in calculus-based introductory physics). A single point on the CRT is awarded for each correct answer, and a score of 2 or 3 is regarded as indicating a strong tendency to mediate intuition with analytical thinking.

3. Interaction between mindware and cognitive reflection

One aspect of the interaction between mindware and the tendency toward cognitive reflection is particularly relevant to our work. Based on the DPTOR framework, there are three possible sources of processing error that lead to an incorrect inference: (i) missing or inadequate mindware, (ii) failure to initiate a conflict detection (i.e., failure to engage in cognitive reflection) between a default (miserly) response and the learned normative mindware, and (iii) failure to sustain an override of the default model with a new model consistent with the normative mindware. The "upstream" process of conflict detection must occur in order for the "downstream" process of a sustained override to take place. The sustained override could be particularly challenging even for reasoners with strong cognitive reflection skills. As stated above, even if a reasoner engages in cognitive reflection and questions the validity of the default model, in many real situations the criteria that need to be satisfied in order to accept or reject the default model are not readily available. For example, a student who has learned Newton's 2nd law at the level adequate to apply it correctly in the basic situation of the block at rest on a table, may not recognize that the same mindware could be used to check the validity of the default response to the magnet question. Instead, process 2 may take a different approach: this student may recognize that the default response that "friction opposes the force of the hand" has worked successfully in solving other physics problems. The student may perceive that a criterion related to how often this approach has been successful is appropriate for establishing the validity of the argument, rather than checking for consistency with the relevant mindware. In a recent paper by Stanovich, he points out that a minimal level of mindware as well as cognitive reflection skills are necessary to initiate the process of conflict detection, but more deeply instantiated mindware is required for a sustained override. Indeed, he notes that "conflict is sometimes detected even in cases where a subject did not successfully override the intuitive response." [47,48]. From our own work, the quote below from a physics student illustrates her reflection on this processing error: "*I tried to formally think it through by trying to remember the magnitude equations and by thinking of the different forces or anything that may be relevant to the problem. I was uncertain with my formal thinking attempt, so I went with my intuitive reasoning*."

E. Summary of diagnostic outcomes from theoretical framework

Dual-process theories suggest that stronger mindware—such as, for example, a conceptual understanding of Newton's laws—impacts physics performance in a variety of ways, but that cognitive reflection is likely to be most relevant in those cases in which a student's first-available mental model is incorrect. Questions developed through physics education research, such as the magnet question, often elicit a strong intuitive response, and thus frequently require successful detection and override to answer correctly. We are particularly concerned about poor student performance on such questions even from students who

successfully applied the relevant mindware in other situations. We argue that by applying the screening-target methodology coupled with the CRT as a measure of the tendency toward cognitive reflection, it is possible to identify a set of diagnostic outcomes that suggest specific approaches for further curriculum development. In particular, if student performance on the screening question is adequate, thus suggesting the presence of (at least) a minimal level of mindware necessary to reach a correct response, then two outcomes are worth considering.

1. Outcome 1

Student performance on the target question is worse than that on the screening question and no relationship between performance on the target question and CRT score is detected. This outcome may occur after basic instruction on relevant topics but before targeted instruction focusing on the inconsistency between students' reasoning on the screening question and their responses to the target question. This outcome suggests that even those students who possess the ability to recognize and mediate instances of intuitive thought (i.e., those who tend toward cognitive reflection as opposed to miserly processing) do not yet possess mindware strong enough either to check for consistency between the normative mindware and the default response or to resolve such inconsistencies. Instruction that focuses on strengthening mindware and building coherence in student knowledge may be necessary in order to initiate and sustain a productive override. In our study, this outcome was observed on a pretest, as discussed in Sec. II B 2. Specific instructional strategies were proposed in Sec. IV and evaluated in Sec. V B.

2. Outcome 2

Student performance on the target question is worse than that on the screening question and a positive relationship is detected between performance on the magnet question and CRT score. This outcome suggests that perhaps students did develop mindware coherent enough to recognize and resolve inconsistencies between normative mindware and the default response, but only students with high CRT scores benefitted by engaging in cognitive reflection and utilizing such mindware productively. In our study, this outcome occurred after our instructional intervention, as discussed in Sec. V B. Our DPTOR-based framework suggests two possible approaches for improving performance even further: (i) focus on improving student cognitive reflection skills, or (ii) improve student mindware to the level of automaticity such that a default response is consistent with the normative mindware independent of cognitive reflection skills. See Sec. IV A for further discussion of the advantages and limitations of these kinds of approaches.

III. PHASE 1: DISENTANGLING THE CONTRIBUTIONS OF MINDWARE AND COGNITIVE REFLECTION SKILLS TO PRODUCTIVE PHYSICS REASONING PRIOR TO THE INTERVENTION

A. Methodology

For simplicity and clarity of the discussion, the main body of this paper describes our investigation conducted in the context of a single student population. The reproducibility of our results from two additional student populations is discussed in Sec. VI.

Data were collected in the introductory calculus-based mechanics course at a research-focused university in the U.S. Students taking the course were primarily intending to pursue majors in the physical sciences or in engineering. The intervention was administered and data were collected as part of the laboratory component of the course.

This project utilized four data streams. Three of these streams involved physics content questions related to the block-magnet sequence shown in Fig. 1, including (i) data collected before any intervention had taken place but after all relevant lecture instruction ("pretest" data); (ii) data that were collected as part of the instructional intervention ("intervention" data), which included three distinct stages described in detail in Sec. IV; and (iii) data that were collected as part of a course exam ("post-test" data). For all data from streams 1–3, responses were coded as either "correct" or "incorrect." A response was coded as correct if it contained the correct answer supported by correct reasoning. The fourth stream was the 3-item CRT described in Sec. II. The CRT was administered early in the term and was included at the end of a regular ungraded "pretest" assignment in advance of a weekly lab. Students received participation credit for completing the CRT, regardless of the correctness of their responses.

We used binary logistic regression to probe the relationships among variables of interest. In a logistic regression analysis, a model is constructed that predicts the probability of a particular dichotomous outcome (here, whether a student will respond correctly or incorrectly to a question) based on the value of various predictor variables (such as a student's score on the CRT or performance on a particular pretest question). A more detailed discussion of logistic regression is included in Appendix.

B. Preintervention assessment: Data, analysis, and results

The pretest was administered online outside of class as part of a regular ungraded assignment in advance of a weekly lab. Students were asked to complete the assignment individually, and credit was given based on effort rather than correctness of responses. Students were shown the block-magnet question pair on a single page of the online assignment. They were asked to answer each

question in a multiple-choice format and were provided with a text box in which to explain their reasoning.

1. Discussion of inconsistencies in student performance

As shown in Table I, before the intervention, approximately two-thirds of the students answered the block question correctly. The overwhelming majority of students who answered the question *incorrectly* apparently did not attempt to apply Newton's 2nd law. The most prevalent incorrect response seems to have been cued by the notion that a massless rod cannot exert a force since, as one student noted, "the mass of the rod is zero, $F = ma$, so the force by the rod must also be zero." A smaller fraction of students used the included figure to argue that "the hand appears to push on the rod, so the rod pushes down on the block." There is no evidence that these students attempted to answer the block question by applying appropriate formal physics knowledge.

Student performance on the magnet question was much weaker than that on the block question. Only ∼20% of the students answered the magnet question correctly. A significant fraction of those students who correctly applied Newton's 2nd law on the block question failed to do so on the magnet question. Less than 30% of the students who answered the block question correctly also applied Newton's 2nd law consistently on the magnet question. As discussed in Sec. II A, the overwhelming majority of the students who answered the magnet question incorrectly argued that "the force of friction must point down because friction opposes the applied force by the hand."

2. Interpretation of inconsistencies through DPTOR

We find that on two questions that both require the application of Newton's 2nd law, student performance differed dramatically. Students did not apply the necessary mindware in a consistent manner. Our theoretical framework posits that people reason incorrectly because (a) they do not possess adequate mindware, or (b) they do not engage in productive cognitive reflection. To determine which instructional strategies would best help students recognize and address inconsistencies in their reasoning, it was necessary to at least attempt to disentangle mindware from cognitive reflection skills and to investigate the relationships among these two factors and student performance on the magnet question.

Identifying relationships among mindware, cognitive reflection skills, and student performance on the pre-intervention magnet question.—Logistic regression models were generated in order to test the relationships among student performance on the magnet question, the presence of at least a minimum level of mindware (as measured by performance on the block question), and the tendency toward cognitive reflection (as measured by the CRT). As discussed earlier, a correct answer on the block question appears to be necessary, but not sufficient, for a correct answer on the magnet question, which is confirmed by the results of the logistic regression models shown in Table II. (Model statistics: $\chi^2 = 11.4$, $p = 0.001$). The CRT score, however, does not appear to be a predictor of success on the magnet question.

Through the lens of DPTOR, the results suggest that the majority of students who acquired the mindware necessary to answer both questions correctly seemed to abandon the correct line of reasoning on the magnet question in favor of a more readily available mental model consistent with the notion that "the force of friction opposes the applied force." This mental model may be more readily available (and therefore intuitively appealing) to the students because it is often used correctly and appropriately in a variety of situations discussed in introductory mechanics courses. For example, if a block remains at rest on a horizontal surface even though it is pushed by a single horizontal force, then it is inferred that a frictional force also acts on the block in a direction opposite to the applied force. In fact, the concepts of static and kinetic forces of friction are commonly introduced in the context of a block, on a horizontal surface, being pushed or pulled by a single horizontal applied force [49–51]. In addition, the notion that "friction must be overcome for an object to move" is consistent with everyday experience and therefore makes intuitive sense as well. It is thus not surprising that this readily available mental model that "friction opposes the applied force by the hand" immediately comes to mind and is perceived as relevant and correct.

The results of the logistic regression model suggest no association between student performance on the magnet question and the level of cognitive reflection skills. This result is consistent with outcome 1 predicted by our

TABLE I. Average CRT scores and performance results from the block and magnet questions administered before the intervention ($N = 91$).

⟨CRT⟩	1.99
Performance on the block question (pre-intervention)	64%
Performance on the magnet question (pre-intervention)	20%
Performance on the magnet question of only those students who answered the block question correctly	$N = 58$ 29%

TABLE II. The results of logistic regression models that link student preintervention performance on the magnet question and the presence of mindware (as measured by the performance on the block question).

Dependent variable	Predictors	Coeff	p	Exp(β)
Performance on the block question preintervention	Intercept	−3.47	0.001	0.03
	Mindware	2.59	0.014	13.3

theoretical framework. Indeed, even those students who do not tend to immediately jump to conclusions and accept a first-available mental model as correct must identify a compelling reason to go against their intuitive responses. After having applied certain ideas repeatedly in many different situations, novice learners may perceive these ideas as generalizable or "an absolute truth." For example, for most students outside of elementary school, the answer to the question "What is 3 + 3?" comes with "cognitive ease" [23]: they have practiced this addition successfully in many other situations and therefore may not perceive the need to check for the correctness of their answer by engaging in unnecessary counting on fingers. Similarly, it is possible that, for some students in our study, the learned heuristic "friction opposes the applied force" reached a certain level of automaticity for the reasons described above: they have applied this model of friction successfully in many other situations and therefore may not have perceived the need to question the applicability of that model in the context of the magnet problem. Moreover, even if a student raised doubts about the applicability of the "friction opposes the applied force" model to the given situation, the normative mindware associated with Newton's 2nd law may not have been strong enough for the student to recognize that Newton's 2nd law must be used as *a criterion* for checking the validity of that mental model. As such, it appears that even students with a high tendency toward cognitive reflection did not successfully engage in the processes of conflict detection and override due either to a strong feeling of rightness that accompanied the heuristic response or to a weakness in normative mindware (or possibly both).

The lack of dependence on the CRT score furthermore suggests that those students who answered the magnet question correctly did so because their first available mental models were likely to have been based on a correct and formal application of Newton's 2nd law. We suspect that the normative mindware of these students was strong enough that they immediately and subconsciously recognized the magnet question as one that is solved by applying Newton's 2nd law; in other words, the mindware of these students had approached the level of automaticity.

The pretest results, interpreted through the lens of DPTOR, suggest that student performance could be improved by reducing the feeling of rightness of the "friction opposes the applied force" response and by helping students recognize that Newton's 2nd law could not only be used as a tool for solving physics questions (e.g., by "balancing forces" on the block question), but also as a criterion that must be satisfied in considering the applicability of a specific heuristic (e.g., if the vector sum of all the forces in the magnet question does not yield zero, the solution is incorrect). We speculated that those students who answered the screening question correctly, but failed to do so on the target question, possessed the first aspect of the necessary mindware, but needed to acquire (or strengthen) the second aspect as well, which is required for a successful conflict detection and override.

IV. PHASE 2: DESIGN OF INTERVENTION: RATIONALE, STRUCTURE, AND METHODOLOGY

On the surface, the instructional goal of our intervention was to improve student performance on questions similar to the magnet question—namely, questions on which many students may not be able to apply the relevant mindware that they may already possess. Our underlying focus was to use DPTOR and the associated cognitive constructs of mindware and cognitive reflection as a framework for doing so. In this section, we review our overall approach and then describe the stages of the intervention, the rationale behind each stage, details of implementation and data collection.

A. Overall rationale

As demonstrated by students' pretest performance, many students appeared to possess the first aspect of the relevant mindware necessary to solve the magnet question correctly, namely they demonstrated the ability to apply Newton's 2nd law to an object at rest in the absence of salient distracting features. However, many of these students failed to apply this mindware on the magnet question due to the presence of an alternative, highly accessible mental model that seems to overshadow the relevance of Newton's 2nd law. The lack of a relationship between student performance on the magnet question and CRT score suggests that many students did not perceive Newton's second law as a criterion that must be satisfied in order to check for the validity of their first-available intuitive responses, thus lacking mindware necessary for conflict detection and override. This suggested a need to design instruction to help students improve both aspects of mindware with a specific focus on strengthening the second aspect: helping students recognize how to apply Newton's 2nd law as a criterion that needs to be satisfied while checking the validity of a particular response (e.g., deciding whether the learned heuristic that friction opposes the applied force holds in a given situation). For this instructional approach, we would anticipate that post-assessment performance would be related to both a student's mindware (i.e., did a student acquire the two aspects of mindware described above?) and a student's cognitive reflection skills (i.e., did a student recognize the need to check the validity of their first-available response by applying an appropriate criterion?).

Alternatively, in order to improve student performance, we could focus exclusively on improving the first aspect of the normative mindware associated with Newton's 2nd law described above to the point of automatic and immediate recognition of the relevance of that mindware to the magnet task [48]. This could be achieved, for example, by training

students to apply Newton's 2nd law in a variety of situations in which an object remains at rest. After such training, Newton's 2nd law would likely become a highly accessible mental model, which most students would automatically apply to a situation analogous to the magnet question. In this case, the high accessibility of this mental model would overshadow all other alternative mental models and would render the engagement of cognitive reflection skills unnecessary for most students. (Indeed, as was the case on the preintervention magnet task, cognitive reflection skills would not be expected to strongly impact student performance since essentially all students giving a correct answer on the postassessment would have extremely robust mindware ensuring that Newton's 2nd law would be a highly accessible model independent of the cognitive reflection tendencies.) However, while this approach would probably improve student performance, it may not help students build coherence and learn how to check and resolve inconsistencies between the learned heuristics and the normative mindware.

A third option, instruction focusing exclusively on the engagement of cognitive reflection skills without paying attention to mindware, is not likely to be fruitful since, according to Stanovich, in the absence of relevant mindware, both the detection of red flags and productive exploration of alternatives are extremely unlikely [25,48]. Indeed, this appeared to be the case on the pretest (see Sec. III B 2). In another project by our research team, we found that an intervention designed to support productive analytic engagement had little to no impact on students who lacked the requisite mindware [30].

An argument could be made that each of the first two approaches outlined above could be appropriate under certain conditions. For example, the second approach may be favored in cases in which an automatic application of certain types of essential mindware (e.g., algebraic or vector operations, unit conversion, or application of the right-hand rule) is critical for successful reasoning and problem solving on more challenging tasks [52]. However, in our context, we argue that it is more pragmatic to focus on designing instructional interventions that foster the development of mindware necessary to support productive cognitive reflection because cognitive reflection is a critical element of reasoning and thus one of our instructional goals for its own sake. As such, we tried to design an intervention that helps students develop or strengthen both aspects of the normative mindware discussed above and supports the engagement of their cognitive reflection skills. The latter could be achieved by providing opportunities to recognize the applicability of Newton's 2nd law as a criterion that needs to be satisfied in checking for consistency between the first available response and the normative mindware. This approach would help students learn to recognize red flags in their reasoning (i.e., engage in conflict detection) and to resolve possible inconsistencies (i.e., execute a sustained override).

We structured the intervention in three stages. In stage 1, students work individually on a task designed to raise awareness of similarities between the block and magnet questions; during this stage, students are given opportunities to recognize that the condition necessary for keeping an object at rest ($\vec{F}_{\text{net}} = 0$) must be satisfied independent of the types of forces involved. In stage 2, students work in groups to analyze the original block and magnet question pair. In stage 3, students work in groups through a more scaffolded activity, answering questions designed to guide them to apply Newton's 2nd law and to refine their ideas surrounding the alternative model "friction opposes the hand." The timing of the pretest, intervention, and post-test relative to instruction on forces is shown in Fig. 4. Below, we describe the rationale for and implementation of each stage in detail.

B. Intervention stages: Overview and rationale

1. Stage 1: Individual work intervention designed to raise awareness of similarities between block and magnet questions

Stage 1 was given immediately after students completed the block-magnet question pair on the pretest as part of an online assignment administered outside of class in advance of a weekly lab or recitation. Students were taken to a new page in which they were provided with a (correct) solution to the block question, shown in Fig. 5, and were asked to indicate whether or not they agree with the given solution and to explain their reasoning. The text of the solution was written using generalized terms such that it did not explicitly reference the physical context (e.g., the block, the specific forces acting on the block) in order to foreground the correct approach to analyzing forces acting on an object at rest. (To be consistent with a force labeling convention practiced in class, the

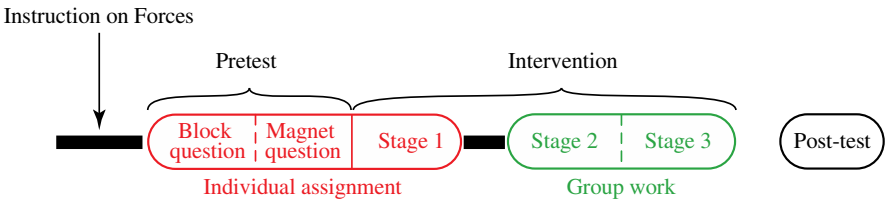


FIG. 4. Instructional sequence.

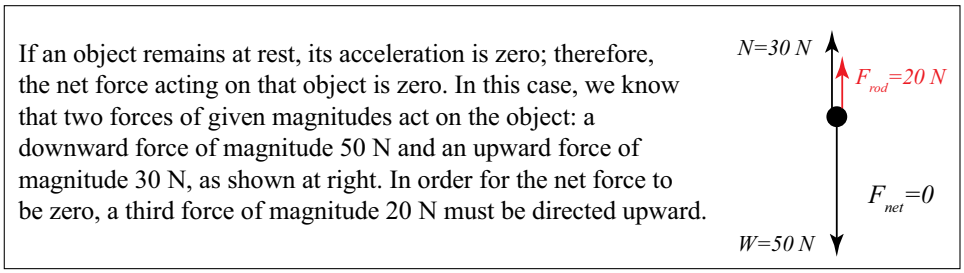


FIG. 5. Feature-free solution to the block question as part of stage 1.

accompanying free-body diagram did reference the specific forces acting on the block, as shown in Fig. 5.) Next, the students were asked whether they still agreed with their original answer to the magnet question, and to explain why they agreed or disagreed. If a student responded that he or she disagreed with his or her original response, the student was given another opportunity to respond to the magnet question, and this response was recorded as that student's stage 1 response.

While it is important to note that this feature-free solution to the block question does not directly connect the block and magnet questions and does not explicitly state that a similar analysis should be used in both cases, it was intended to act as a gentle nudge for students, potentially triggering a more productive engagement of the analytic process by raising awareness of similarities between the two scenarios. According to DPTOR, if students recognize similarities between both scenarios while noting their discrepant reasoning approaches, it is more likely that potential red flags associated with their first-available mental models for the magnet question will impact their reasoning process (e.g., by lowering their feeling of rightness and fostering a deeper engagement of the analytic process). This stage of the intervention, therefore, had the potential to help students (i) recognize similarities between the two scenarios (i.e., both objects are at rest), (ii) recognize that conditions for an object to remain at rest ($\vec{F}_{\text{net}} = 0$) must be satisfied independent of the specific features of a given situation (e.g., types of forces and types of objects), and (iii) identify and resolve inconsistencies between the intuitive responses to the magnet question and the normative mindware.

2. Stages 2 and 3: Group work

Students worked in small groups in stages 2 and 3 of the intervention sequence. The group work format was expressly chosen since it allows for socially mediated metacognition [36,53], in which the group members help shape and guide the thinking and reasoning of the group. The group, which is the effective unit of analysis, necessarily engages in self-assessment and self-regulation and draws upon its collective metacognitive knowledge while working collaboratively. The metacognition of the group is effectively externalized; with members generating new ideas, assessing information and approaches, disclosing their own thinking, requesting feedback on their own thinking, and monitoring their partners' thinking [53]. Through this collaborative thinking process, it is more likely that red flags will be identified and that intuitive ideas may be mediated via analytical reasoning—namely, collaborative thinking has the potential to support the cognitive reflection of the group. In addition, by externalizing metacognition and the nature of reasoning, it is plausible that group work may help students learn the value of and strategies for cognitive reflection, which can in turn be employed while working individually.

Stage 2: Group work intervention involving the block and magnet pair.—This stage, which enabled students to revisit the block-magnet question pair in groups, was administered in the laboratory component of the course a short time after students completed the web-based stage 1. Students worked with their regular lab partners in groups of two. (A few groups had three students.) Each group was tasked with discussing their approaches to the questions until a group consensus was reached, at which point a single group consensus response (including both an answer and reasoning) was submitted via a web-based form for each of the two questions. As discussed earlier, this intervention was designed to foster cognitive reflection and promote consistency checking via socially mediated metacognition. In addition, we anticipated that the opportunity to collaboratively revisit the block and magnet question pair could also help strengthen student mindware, if necessary.

Stage 3: Group work intervention consisting of a sequence of guiding questions.—This stage of the intervention was designed exclusively for those lab groups that did not answer the magnet question correctly after stage 2. Using the same web-based online system, these groups were automatically served a final sequence of questions that was intended to provide more scaffolding and step-by-step guidance for analyzing the magnet question. In the sequence of guiding questions, groups were shown the magnet question once again, but were asked to consider two different scenarios. In the first scenario, the hand is not exerting a force on the stationary magnet. The second scenario is identical to the original magnet question, in which the hand *is* applying a force of 6 N upward on the magnet. Students

were asked to draw a free-body diagram for the stationary magnet in both cases; they were then asked to determine the net force on the magnet in each case and to record their reasoning. Students were also asked what their responses suggest about the direction of the forces of friction acting on the magnet in both scenarios and to explain their reasoning. Students were expected to recognize that Newton's 2nd law must hold in both scenarios; however, in the absence of a hand, the friction *does* oppose the applied force, which, in this case, is the force of gravity (and not that applied by the hand), while in the presence of the hand, the force of static friction must oppose the vector sum of the forces by gravity and the hand. Therefore, the activities of Stage 3 emphasized that in some cases the heuristic response of "friction opposing the applied force" is consistent with the normative response while in others a more careful analysis via the application of the Newton's 2nd law is required. Hence, stage 3 was designed to stress that Newton's 2nd law must be satisfied in validating a specific response.

C. Intervention data and analysis strategies

Student responses to the intervention questions were coded as either "correct" or "incorrect." A response was coded as correct if it contained the correct answer supported by the correct reasoning. For stages 2 and 3, a correct code assigned to a student indicates that the student had been part of a group that had responded correctly, suggesting that the student agreed with the correct answer even if he or she did not necessarily generate it on their own.

We also created a variable to measure whether a student had ever responded correctly to the magnet question during any intervention stage. This variable, called Any_point_magnet, was coded as a 1 if any of the responses to the magnet question, either on the pretest or during any stage of the intervention, was correct. If a student had never provided (or been part of a group that submitted) a correct response to the magnet question, this variable was coded as a 0.

Student performance on the pre-intervention block question does not serve as a suitable measure of mindware in relation to the post-test performance because improved student performance at every intervention stage suggests improved mindware as a result of the intervention. In this study, we argue that the variable Any_point_magnet could serve as a reasonable estimate of the presence of mindware after the intervention in the sense that a student's successful performance on the magnet question during the intervention could serve as an indication that the student either successfully applied the normative mindware to arrive at the correct answer or successfully applied the normative mindware in order to detect and override an intuitively appealing response. In either case, we argue that the student possessed and successfully practiced applying the mindware necessary to answer the magnet question correctly or, at the very least, the student participated in a discussion with a lab partner and agreed with the correct solution that contradicted his or her individual response. We note that there are several possible ways to use the data collected in this study to estimate mindware (including, for example, taking into account whether a correct response was given during only the individual component of the intervention). We found that the results were largely consistent and independent of a specific approach. As such, we adopted the most parsimonious model in which the variable Any_point_magnet represents a reasonable estimate of the presence of mindware. We also note that the variable Any_point_magnet provides the most proximal (in content) measure of mindware. Given that we are interested in probing the extent to which student post-test performance is linked to cognitive reflection skills (in addition to mindware), the detection of such a link even after the most proximal measure of mindware is taken into account strengthens any claims that we may make. Therefore, we argue that the endorsement of a correct answer to the magnet question at any point during the instructional sequence is an adequate (most parsimonious and proximal) measure of the presence of relevant mindware.

V. PHASE 3: USING ASSESSMENT DATA TO EXAMINE IMPACT OF INTERVENTION ON ASPECTS OF PRODUCTIVE REASONING

A. Postintervention data collection

On the intervention post-test, a four-part question in the context of a stationary magnet on a refrigerator was administered as part of a course exam (shown in Fig. 6). Students were asked to identify the direction of the static friction force on the magnet in each of the four cases. The correct answers for Cases 1–4 are that the friction forces are upward, zero, upward, and upward, respectively. The post-test was administered in a free-response format and included on a standard midterm exam. Responses were coded as correct if the student indicated the correct direction for the friction force in *all four* of the cases and supported these answers with correct reasoning.

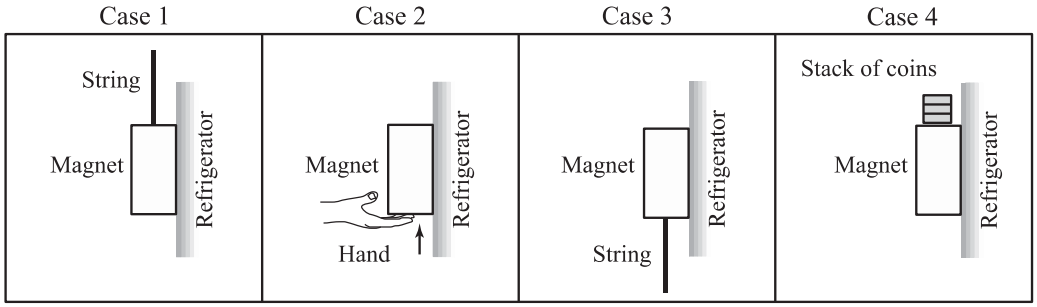
B. Mid- and postintervention assessments: Data analysis and results

As shown in Table III, every intervention stage improved student performance by ∼10%–25%. No single intervention stage was *clearly* more impactful than the rest. After the set of three interventions, the majority of the students (∼75%) gave a correct response to the magnet question at least once at some point during instruction (or at least agreed with the correct answer).

The data indicate that there was a dramatic increase in student performance from pre- to post-test (from ∼20% correct on the magnet question to ∼60% correct on the post-test). However, these results are somewhat less impressive once two factors are taken into account. First, most students in the study had at least three opportunities to

In the four cases below, four identical magnets are attached to the same refrigerator. Each magnet weighs 5 N. Each of the four magnets is observed to remain at rest.

- In Case 1, a string with 3 N of tension pulls *upward* on the magnet.
- In Case 2, a hand pushes *upward* with 5 N of force.
- In Case 3, a string with 3 N of tension pulls *downward.*
- In Case 4, a stack of gold coins that weighs 3 N sits on the top of the magnet.



For all Cases 1-4, magnet weight = 5 N and all magnets remain at rest.

For each of the cases, state whether the static friction force exerted on the magnet by the refrigerator is *upward, downward,* or *equal to zero.* Explain. If there is not enough information to determine this, or if there is more than one direction of static friction in a given case, state that explicitly.

FIG. 6. The post-test question.

consider the magnet question and to modify their thinking. Second, roughly 30% of those students who gave (or at least agreed with) a correct answer on the magnet question at any point during the intervention reverted back to the incorrect notion that friction opposes the applied force by a singular and highly salient external agent. Our results indicate that this identified pattern of student reasoning is very robust, persistent, and does not seem to be easily altered by the "quick fixes" that each interventional stage was designed to afford.

Further data analysis was conducted in order to pinpoint more precisely why, even after targeted instruction specifically designed to help students recognize and resolve inconsistencies between their intuitive responses and normative mindware, some students were successful on the post-test while others were not. Results of a logistic regression model, shown in Eq. (1), suggest that both mindware and the tendency toward cognitive reflection play a role in a student's performance on the post test. The logistic regression model predicts the probability of success on the post-test as a function of two variables: (i) whether a student had ever responded correctly to the magnet question during any of the intervention stages (called Any_point_magnet and taken as a measure of mindware as discussed in Sec. IV C), and (ii) cognitive reflection skills, as measured by CRT score. (All coefficients are statistically significant at the $p < 0.05$ level, model statistics are $\chi^2 = 12.2$, $df = 2$, $p < 0.01$).

$$p = \frac{1}{1 + e^{(1.7 - 1.4 \times \text{Any\_point\_magnet} - 0.6 \times \text{CRT})}}. \qquad (1)$$

TABLE III. Student performance on pretest, intervention questions, and post-test.

Performance on the block question (pretest)		64%
Performance on the magnet question	Pretest	20%
	Stage 1	32%
	Stage 2	49%
	Stage 3	73%
	Correct at any point during the intervention process	74%
Correct on the post-test		62%
Correct on the post-test of those who were correct at some point during instruction		69%

The results suggest that those students who gave a correct response to the magnet question at least once during the intervention and who also possessed high cognitive reflection skills were much more likely to answer the post-test question correctly compared to those who answered the magnet question correctly during the intervention but scored zero on the CRT. For example, the odds of success on the post-test question increases by a factor of approximately 6 (given by $e^{0.6\times3}$) for a student who answered correctly during instruction and received a score of 3 on the CRT compared to a student who also answered the magnet question correctly during instruction but scored zero on the CRT.

The results suggest that the success of our intervention was mediated by the students' cognitive reflection skills. Indeed, students with a stronger tendency toward cognitive reflection seemed to access the necessary mindware successfully under exam conditions. These students either immediately recognized the applicability of Newton's 2nd Law to the post-test or were able to engage in successful cognitive reflection. The latter would first involve the initiation of an "upstream" process of conflict detection (e.g., asking whether the heuristic that friction opposes the applied force by the string is applicable here) followed by the "downstream" process of a sustained override that utilizes Newton's 2nd Law as a criterion for checking the validity of the heuristic response. However, even in the presence of the mindware, our analysis indicates that students with weaker cognitive reflection skills tended to jump to conclusions and accept the intuitively appealing response that friction opposes the applied force by a highly salient agent (e.g., a hand or a string,) as correct. Through the lens of DPTOR, this result suggests that these students most likely failed to initiate the "upstream" process of conflict detection and therefore never had the opportunity to utilize their mindware for sustaining a productive override. For such students, the impact of the intervention appeared to be short lived. These results are consistent with outcome 2 discussed in Sec. II E.

VI. REPLICABILITY OF RESULTS

We were particularly interested in exploring the extent to which the results from our intervention and analysis could be replicated at different institutions. If results were similar across multiple institutions, it would be more likely that the proposed mechanisms were, in fact, at play and that our findings could be more easily generalized. We therefore administered the intervention in the introductory calculus-based mechanics course at two additional universities (B and C in Fig. 7) in the U.S., which serve a diverse range of students as measured by incoming Math SAT scores. (Note that University A corresponds to the primary implementation site.)

The intervention implementations at these other sites necessarily varied somewhat due to instructional constraints. At University B, the intervention was administered and data were collected as part of a required laboratory component of the mechanics course, as at University A. Multiple lecture sections, taught by different instructors, fed into the laboratory. At University C, the intervention was administered via a sequence of clicker questions and data were collected from three lecture sections of the course, taught by three different instructors in the same academic term. In two of those sections, administration of the intervention was identical to one another; therefore, data from those sections have been combined (referred to as Section I). In the remaining section, referred to as Section II, students were only given the stage 1 intervention

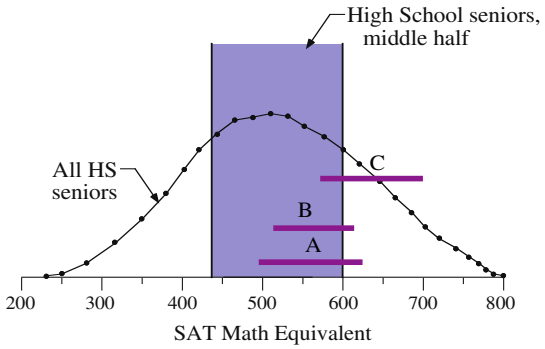


FIG. 7. 25th–75th percentile of incoming math SAT scores from the three populations, relative to national values. University A is the primary implementation site.

prior to the post-test. In all courses, instructors used research-based, active learning strategies to various degrees. The sample sizes from University B, University C (Section I), and University C (Section 2) were $N = 125$, $N = 230$, and $N = 104$, respectively.

Prior to the intervention, logistic regression of data from Universities B and C revealed that a correct response to the block question was necessary, but not sufficient, for a correct response to the magnet question. CRT score, however, was not a predictor for success on the magnet question. Thus, at all three universities, outcome 1 (predicted by our theoretical framework) was documented prior to the intervention, suggesting that, at this stage, students did not yet possess mindware that was sufficiently robust to support consistency checking between their default response and the normative mindware or to resolve such inconsistencies.

After the complete intervention (University B and University C, Section 1), logistic regression models indicated that the probability of success on the post-test depended on both mindware (as measured by the variable Any_point_magnet) and the tendency toward cognitive reflection (as measured by the CRT); these findings were thus consistent with those from the primary implementation site (University A), and were illustrative of outcome 2 predicted by our theoretical framework. Our results suggest that the full intervention sequence supported students in developing sufficiently robust mindware to help them recognize and resolve the inconsistency between their default response and the normative mindware—provided they have a stronger tendency toward cognitive reflection (i.e., a high CRT score).

It is particularly important to note that no such dependence on CRT score was observed for students from University C Section 2, who only participated in stage 1 of the intervention (not stages 2 and 3). For these students, performance on the post-test was much weaker than that observed with the other student populations in this study (in fact, no significant improvement in performance was observed between the intervention stage 1 and the

post-test). More importantly, the post-test performance of these students was solely linked to mindware (as was observed for *all* student populations on the pretest), suggesting that, in the absence of the complete intervention sequence, many students did not develop sufficiently robust mindware. Indeed, since stages 2 and 3 were not part of instruction, the second aspect of the necessary mindware, which is related to Newton's 2nd law as a criterion for checking for consistency, was not addressed. While these students did apply Newton's 2nd law in a variety of situations during course instruction, without explicit opportunities to recognize and practice the application of Newton's 2nd law for consistency checking (i.e., as a criterion against which to check their default model), these students continued to misapply a learned heuristic even after all instruction was complete.

Collectively, these results reveal that the phenomena we documented at the primary investigation site (University A) are, in fact, more universal and are not idiosyncratic to one particular population or experiment. Our findings also underscore the importance of attending to the development of coherent mindware that enables students to both (i) apply this mindware in order to build a correct model or argument and (ii) apply this mindware strategically for consistency checking. The latter plays an integral role in the conflict detection and sustained override associated with cognitive reflection, which is why post-test performance was observed to depend on *both* mindware and tendency toward cognitive reflection whereas pretest performance depended solely on mindware.

VII. DISCUSSION AND IMPLICATIONS FOR RESEARCH-BASED CURRICULUM DEVELOPMENT

In this investigation, the data collection, analysis, and interpretation were all driven by the idea that two factors, mindware and cognitive reflection, impact student reasoning on physics questions. The design of the intervention sequence was informed by this idea, which is deeply rooted in DPTOR. Indeed, the dual-process framework suggests that in order to switch from a first-available, highly compelling intuitive model to an alternative model based on formal knowledge, students must be able to engage in cognitive reflection so that they may detect the conflict and successfully mediate the intuitive thinking with analytical thinking that draws upon relevant mindware [25,48]. Our empirical study provides evidence that both factors play a critical role in student reasoning on a question that tends to elicit strong, intuitively appealing incorrect responses. Moreover, the relationships among student performance, mindware, and cognitive reflection skills documented for a single student population was similarly observed for two other student populations as well.

A. Summary of findings from phase 1 analysis

Students' pretest performance on the magnet question differed drastically from that on the block question. We establish that it is reasonable to treat the block question as an adequate measure of whether or not a student possesses at least the minimum level of mindware necessary to answer the magnet question correctly (i.e., possessed the first aspect of mindware as discussed in Sec. III B 2). Namely, they were able to apply Newton's 2nd law to balance forces acting on an object at rest. The results of the logistic regression model suggested that while the presence of this mindware was linked to performance on the magnet question, cognitive reflection skills did not appear to be a predictor of success on that particular question. Interpreting the results through the lens of DPTOR, we argued that the level of mindware of those students who answered both questions correctly was likely to be quite strong (i.e., reached the level of automaticity) so that the magnet question cued a first available mental model consistent with a normative response. Thus, in the absence of a conflict, the engagement of cognitive reflection skills was not necessary. Those students who answered the block question correctly, but failed to apply the same mindware on the magnet question, did appear to possess the necessary mindware; however, the strength of that mindware may not have been sufficient to cue the normative response. Instead, for these students, the first available mental model (i.e., friction opposes the force by the hand) appeared to hold strong enough intuitive appeal so that the students either did not perceive the need to question its validity or they missed the second aspect of mindware that would allow them to recognize that Newton's 2nd law must be used as a criterion to be satisfied in order to check for validity of the first-available intuitive response. The engagement of the cognitive reflection skills of these students, therefore, also did not appear to be relevant and predictive of success on the magnet pretest. These results are consistent with outcome 1 for the interaction between mindware and cognitive reflection predicted by DPTOR.

B. Summary of findings from phase 3 analysis

After the instructional intervention sequence, student performance on the post-test revealed significant improvements. Data suggested that the intervention did appear to engage some students in productive cognitive reflection. The logistic regression model indicates that, after the intervention, student performance on the post-test appears to be linked to both success on the magnet question prior to the post-test (e.g., during intervention) and cognitive reflection skills. Indeed, even after controlling for mindware (as indicated by the prior success on the magnet question), students with a higher level of cognitive reflection skills were more likely to answer the post-test question correctly. These results are consistent with outcome 2 for the interaction between mindware and cognitive reflection suggested by DPTOR.

C. Limitations of findings and need for further research

It is important to note that this in-depth investigation, though detailed in design and analysis, was conducted in the context of a single screening-target question pair, intervention, and closely related post-test involving the application of Newton's 2nd law to stationary objects (i.e., near transfer). We argue, however, that the design and refinement of instructional interventions could benefit from the kind of diagnostic assessments made possible by applying the screening-target methodology coupled with the CRT as a measure of the tendency toward cognitive reflection. (See Sec. II E) This particular study presents a case of pre-intervention results consistent with diagnostic outcome 1 only. In our ongoing work, we have also identified sets of screening-target pairs that, coupled with the CRT, yield preintervention results consistent with outcome 2 as well. While we briefly suggest specific instructional approaches (informed by DPTOR) for such cases in Sec. II E, an in-depth discussion of designing, testing, and assessing efficacy of such approaches will be the focus of future papers. In addition, given that the reasoning phenomenon discussed in this paper is neither limited nor unique to the topics of Newton's 2nd law and friction, we wish to probe whether the single intervention presented here (or analogous interventions) will help students successfully employ conflict detection in other, farther removed contexts as well (i.e., far transfer). Indeed, this is the focus of another ongoing investigation.

D. Insights into the nature of student reasoning in physics and implications for physics instruction

On the basis of this investigation, several important insights have emerged related to understanding, interpreting, and supporting student reasoning in the context of physics instruction.

Intuitive processing is deeply ingrained.—Quick and isolated fixes are not likely to produce a desirable impact on reasoning in contexts that elicit strong, intuitively appealing responses. Our analysis suggests that no single intervention stage was more impactful than the rest. In fact, each stage produced a relatively minor improvement in student performance. Moreover, even after the completion of the entire intervention sequence, of those students who gave (or at least agreed with) the correct answer to the magnet questions at some point during instruction, a significant fraction of the students (approximately one-third) reverted back to the incorrect, intuitive reasoning on the post-test. This finding highlights the deeply ingrained nature of intuitive processing. Perhaps the most powerful aspects of DPTOR are the assertions that (i) the first available mental model serves as an entry point into any reasoning path and (ii) process 1 *cannot* be turned off.

Need for instructional focus on cognitive reflection.—Based on our findings, novices may not spontaneously engage in cognitive reflection on physics questions. On the preintervention magnet question, the lack of dependence of student performance on cognitive reflection skills (a similar result was observed on the post-test for those students who did not complete all stages of the intervention) suggests that many students were not yet able to recognize or act upon red flags in their reasoning, even in the presence of the mindware required to solve the magnet problem correctly. As such, interventions designed to help students recognize instances of intuitive thought and to help them develop productive approaches for mediating such thoughts are necessary.

Need for instruction that strengthens both aspects of mindware in order to support productive cognitive reflection and successful student reasoning.—Our results provide evidence that instruction with a primary focus on improving one aspect of mindware, namely how to use it as a tool for solving a specific type of problems (e.g., balancing forces for an object at rest), may not provide optimal opportunities for developing productive reasoning habits in physics and beyond. The inconsistency in student reasoning on the block-magnet pretest (as well as post-test performance of students in University C Section 2) suggests that the presence of this type of mindware alone is not enough to enable students to detect red flags in their reasoning and to mediate intuitive responses with the formal application of necessary mindware. While domain-specific expertise is necessarily multifaceted, the ability to detect and mediate intuition-based responses is an important ingredient necessary for developing expertise. As such, designing curriculum without attending to the second aspect of mindware, namely, how to use it as a tool for productive conflict detection and override, is likely to be less effective at helping students reason successfully on many kinds of physics questions. In essence, attending to the first aspect of mindware alone may promote the development of heuristic reasoning (e.g., "balance forces for an object at rest," or "friction opposes the force by the hand") as opposed to helping students improve their analytical reasoning skills (e.g., checking for consistency and validating their responses).

How the deeply ingrained nature of intuitive processing impacts and informs instruction.—The complex interplay between mindware and cognitive reflection skills may also explain why it is fairly common for students who perform well in classroom activities (e.g., clicker questions, group work) to be less successful on nearly identical test problems. The decline in student performance could be very frustrating to both the instructor and the students. However, from our standpoint, it is an expected, unavoidable, and natural part of developing expertise in physics. One could, in principle, argue that such a decline may be due to suboptimal classroom instruction (or even that "students are not trying hard enough"). However, according to Stanovich, unless mindware has "been practiced to

automaticity" such that it "can automatically compete with (and often immediately defeat) any alternative nonnormative response," there is always room for error in applying that mindware for a student who has not yet developed the habits of mind to recognize instances of intuitive thought and/or to trust formal knowledge enough to override a strong intuitively appealing response [48].

In addition, experienced instructors often hear frustrated students sharing that they studied hard and knew what concepts were relevant to each exam question, but that they frequently second guessed themselves (in the *wrong direction*). We argue that the phenomenon of second guessing, while frustrating to students, is a natural aspect of learning how to reason productively. Second guessing suggests that a student possesses an adequate level of mindware and cognitive reflection skills to consider alternatives; however, the student has not yet developed the habits of mind to either trust their formal knowledge or to successfully analyze and reconcile inconsistencies between the alternative approaches. Most reasoners who experienced the dilemma of "trusting or going against their guts" may attest to the angst that such a dilemma presents. To highlight this feeling of angst, we present a quote from a student who answered a question (similar to the magnet question used in this study) correctly after deciding against his initial intuitive response. This student offered the following reflection: "*I applied the formal reasoning to the best of my ability, unless I thought myself into a wrong answer, which would look idiotic*."

We speculate that physics instruction that makes the dual nature of human cognition explicit and visible to students may impact student learning of physics in ways that extend beyond improvements in student performance. For example, it may help establish an instructional environment that emphasizes that the careful examination (and possible rejection) of an intuitive response is a natural part of the reasoning and *not* an indication of a lack of knowledge or of any other deficiency or failure on the part of the student. Such a reframing may also help address a common concern that research-based materials that elicit incorrect intuitive responses as a part of the instructional cycle may lead to feelings of student inadequacy. Explicit discussion that such incorrect responses stem from the dual nature of human thinking because intuitive processing cannot be turned off may help alleviate such concerns. Developing an awareness of one's own thinking paths as well as the ability to recognize red flags is a critical step toward the development of expertise in physics, and students must be supported throughout the instructional process to strengthen these metacognitive skills.

VIII. CONCLUSION

In this investigation, we developed, implemented, and assessed, at three different institutions, the impact of a single intervention sequence focused on the application of Newton's 2nd law to objects at rest. A family of theoretical frameworks from cognitive science known as dual-process theories of reasoning and decision-making informed the design, assessment, and analysis efforts for the intervention sequence. Indeed, drawing upon dual-process theories, we created the intervention sequence with the aim of supporting productive reasoning by strengthening mindware and promoting cognitive reflection. Data analysis indicated that the intervention improved students' mindware to the level needed for consistency checking, an integral part of cognitive reflection. Moreover, as expected, students' cognitive reflection skills were shown to mediate the success of the intervention, with those students who demonstrated a stronger tendency toward cognitive reflection being more able to access the necessary mindware successfully. Our findings suggest that interventions expressly designed to support the development of mindware such that it may be used as a criterion during cognitive reflection can help students more productively and purposefully mediate intuitive thinking (which cannot be turned off) by reasoning more analytically. Indeed, our analysis has shown, consistent with the findings from other studies, that these kinds of intuitive processing errors may occur despite strong content understanding. We argue that it is important to help students recognize that revisiting one's intuitive thoughts with a critical eye is an integral component of scientific thinking and is necessitated by the way in which the human brain functions. Students should thus be encouraged to view such errors as an inherent part of the thinking process and to recognize that mindware may also be used to test intuitive thinking. It is important to stress that we advocate on behalf of engaging in cognitive reflection—not simply suppressing one's intuition. We are confident that the coherent, theory-driven approach to intervention design and testing illustrated in this manuscript is an important contribution to ongoing efforts to account for and leverage domain-general reasoning phenomena in the learning and teaching of physics.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. DUE-1821390, No. DUE-1821123, No. DUE-1821400, No. DUE-1821511, No. DUE-1821561, No. DUE-1431940, No. DUE-1431541, No. DUE-1431857, No. DUE-1432052, and No. DUE-1432765.

APPENDIX: OVERVIEW OF LOGISTIC REGRESSIONS

We used binary logistic regression to probe the relationships among variables of interest. In a logistic regression analysis, a model is constructed that predicts the probability of a particular dichotomous outcome (in this study, whether a student will respond correctly or incorrectly to a question)

based on the value of various predictor variables (such as a student's score on the CRT or performance on a particular pretest question). The predictor variables can be either categorical or continuous. The logistic regression algorithm fits a multiple linear regression algorithm for the log of the odds of an event:

$$\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n,$$

where x_n are the predictor variables and β_n are regression coefficients estimated by the algorithm. The odds of an event are given by

$$\text{odds} = \frac{P(Y)}{1 - P(Y)} = \frac{\text{probability of event } Y \text{ occuring}}{\text{probability of event } Y \text{ not occuring}}.$$

It can be shown from the two equations above that the probability $P(Y)$ of event Y occurring is then given by

$$P(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}.$$

For each β_n, a Wald statistic is also calculated, which has a chi-squared distribution and allows for a determination of whether the coefficient differs significantly from zero. If the coefficient for a particular predictor is significantly different from zero, that predictor is considered to make a significant contribution to the probability of the event occurring. Logistic regression can thus be used to determine which predictor variables are contributing significantly to a particular outcome.

The odds ratio, given by $\exp(\beta_n)$, is often employed to estimate effect size in the context of logistic regression. It indicates the change in odds resulting from a unit change in the predictor.

The assumptions of binary logistic regression are less stringent than those for linear regression and include assumptions of (i) linearity, (ii) independence of errors, and (iii) a lack of multicollinearity. The first assumption states that for any continuous predictor, there must be a linear relationship between this predictor and the log(odds) of the outcome variable. The second assumption means that separate cases of data should not be related—in other words, no single individual should appear multiple times in the dataset (e.g., at different points in time). The third assumption means that predictors should not be too highly correlated with one another. In our analysis, all assumptions were met satisfactorily.

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

[2] P. W. Laws, *Workshop Physics Activity Guide, Core Volume with Module 1: Mechanics I* (John Wiley & Sons, Hoboken, NJ, 2004).

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

[4] E. E. Prather, T. F. Slater, J. P. Adams, and G. Brissenden, *Lecture-Tutorials in Introductory Astronomy*, 3rd ed. (Pearson, London, UK, 2012).

[5] G. Novak, A. Gavrin, W. Christian, and E. Patterson, *Just-in-Time Teaching: Blending Active Learning with Web Technology* (Addison-Wesley, Boston, MA, 1999).

[6] R. E. Scherr and A. Elby, Enabling informed adaptation of reformed instructional materials, AIP Conf. Proc. **883**, 46 (2007).

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

[8] Paradigms in Physics, http://physics.oregonstate.edu/portfolioswiki/.

[9] D. R. Sokoloff and R. K. Thornton, Using interactive lecture demonstrations to create an active learning environment, Phys. Teach. **35**, 340 (1997).

[10] L. C. McDermott and P. S. Shaffer, Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding, Am. J. Phys. **60**, 994 (1992).

[11] F. Goldberg and S. Bendall, Making the invisible visible: A teaching/learning environment that builds on a new view of the physics learner, Am. J. Phys. **63**, 978 (1995).

[12] M. D. Sharma, I. D. Johnston, H. Johnston, K. Varvell, G. Robertson, A. Hopkins, C. Stewart, I. Cooper, and R. Thornton, Use of interactive lecture demonstrations: A ten year study, Phys. Rev. ST Phys. Educ. Res. **6**, 020119 (2010).

[13] M. Kryjevskaia, A. Boudreaux, and D. Heins, Assessing the flexibility of research-based instructional strategies: Implementing tutorials in introductory physics in the lecture environment, Am. J. Phys. **82**, 238 (2014).

[14] T. I. Smith and M. C. Wittmann, Comparing three methods for teaching Newton's third law, Phys. Rev. ST Phys. Educ. Res. **3**, 020105 (2007).

[15] V. K. Otero and K. E. Gray, Attitudinal gains across multiple universities using the Physics and Everyday Thinking curriculum, Phys. Rev. ST Phys. Educ. Res. **4**, 020104 (2008).

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

[17] A. F. Heckler, The ubiquitous patterns of incorrect answers to science questions: The role of automatic, bottom-up processes, Psychol. Learn. Motiv. **55**, 227 (2011).
[18] 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).
[19] C. Singh, Assessing student expertise in introductory physics with isomorphic problems. I. Performance on nonintuitive problem pair from introductory physics, Phys. Rev. ST Phys. Educ. Res. **4**, 010104 (2008).
[20] A. Elby, What students' learning of representations tells us about constructivism, J. Math. Behav. **19**, 481 (2000).
[21] M. Kryjevskaia and M. R. Stetzer, Examining inconsistencies in student reasoning approaches, AIP Conf. Proc. **1513**, 226 (2013).
[22] B. Frank, S. E. Kanim, and L. S. Gomez, Accounting for variability in student responses to motion questions, Phys. Rev. ST Phys. Educ. Res. **4**, 020102 (2008).
[23] D. Kahneman, *Thinking, Fast and Slow* (Farrar, Strauss, & Giroux, New York, 2011).
[24] J. S. B. T. Evans, The heuristic-analytic theory of reasoning: extension and evaluation, Psychon. Bull. Rev. **13**, 378 (2006).
[25] K. E. Stanovich, *What Intelligence Tests Miss: The Psychology of Rational Thought* (Yale University Press, New Haven, CT, 2009).
[26] S. Frederick, Cognitive reflection and decision making, J. Econ. Perspect. **19**, 25 (2005).
[27] 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).
[28] M. Kryjevskaia, Examining the relationships among intuition, reasoning, and conceptual understanding in physics, in *Upgrading Physics Education to Meet the Needs of Society*, edited by M. Pietrocola (Springer, New York, NY, 2019), pp. 181–188, ISBN 978-3-319-96162-0.
[29] C. R. Gette, M. Kryjevskaia, M. R. Stetzer, and P. R. L. Heron, Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy, Phys. Rev. Phys. Educ. Res. **14**, 010113 (2018).
[30] J. C. Speirs, M. R. Stetzer, B. A. Lindsey, and M. Kryjevskaia, Leveraging dual-process theories of reasoning to explore and support student reasoning in physics via reasoning chain construction tasks, Phys. Rev. Phys. Educ. Res. (to be published).
[31] C. Singh, When physical intuition fails, Am. J. Phys. **70**, 1103 (2002).
[32] H. A. Simon, What is an explanation of behavior?, Psychol. Sci. **3**, 150 (1992).
[33] M. Kryjevskaia, M. R. Stetzer, and T. K. Le, 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*, edited by P. V. Engelhardt, A. D. Churukian, and D. L. Jones (AIP, New York, 2015), p. 143.
[34] S. Mamede, T. A. W. Splinter, T. Van Gog, R. M. J. P. Rikers, and H. G. Schmidt, Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes, BMJ Qual. Saf. **21**, 295 (2012).
[35] M. Osman and R. Stavy, Development of intuitive rules: Evaluating the application of the dual-system framework to understanding children's intuitive reasoning, Psychon. Bull. Rev. **13**, 935 (2006).
[36] V. A. Thompson, Dual-process theories: A metacognitive perspective, in, *Two Minds Dual Process. Beyond*, edited by K. Frankish and J. S. B. T. Evans (Oxford University Press, Inc., Oxford, UK, 2012), pp. 171–195.
[37] V. A. Thompson, J. A. Prowse Turner, and G. Pennycook, Intuition, reason, and metacognition, Cogn. Psychol. **63**, 107 (2011).
[38] V. A. Thompson, J. S. B. T. Evans, and J. I. D. Campbell, Matching bias on the selection task: It's fast and feels good, Think. Reas. **19**, 431 (2013).
[39] P. N. Johnson-Laird, *How We Reason* (Oxford University Press, Inc., New York, 2006).
[40] M. E. Toplak, R. F. West, and K. E. Stanovich, The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks, Mem. Cogn. **39**, 1275 (2011).
[41] R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises, Rev. Gen. Psychol. **2**, 175 (1998).
[42] S. Tishman, E. Jay, and D. N. Perkins, Teaching thinking dispositions: From transmission to enculturation, Theory Into Practice **32**, 147 (1993).
[43] G. Pennycook, J. A. Cheyne, D. J. Koehler, and J. A. Fugelsang, Is the cognitive reflection test a measure of both reflection and intuition?, Behav. Res. Meth. Instrum. Comput. **48**, 341 (2016).
[44] G. Campitelli and P. Gerrans, Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach, Mem. Cogn. **42**, 434 (2014).
[45] E. J. N. Stupple, M. Gale, C. R. Richmond, K. Road, D. De, M. Gale, and C. R. Richmond, Working memory, cognitive miserliness and logic as predictors of performance on the cognitive reflection test, in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, edited by M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Cognitive Science Society, 2013), p. 1396.
[46] M. E. Toplak, R. F. West, and K. E. Stanovich, Assessing miserly information processing: An expansion of the Cognitive Reflection Test, Think. Reas. **20**, 147 (2014).
[47] W. De Neys, Conflict detection, dual processes, and logical intuitions: Some clarifications, Think. Reas. **20**, 169 (2014).
[48] K. E. Stanovich, Miserliness in human cognition: The interaction of detection, override, and mindware, Think. Reas. **24**, 423 (2018).
[49] K. Perkins, W. Adams, M. Dubson, N. Finkelstein, S. Reid, C. Wieman, and R. LeMaster, PhET: Interactive simulations for teaching and learning physics, Phys. Teach. **44**, 18 (2006).
[50] P. A. Tipler and G. Mosca, *Physics for Scientists and Engineers*, 6th ed. (W.H. Freeman, New York, NY, 2007), Vol. 1.
[51] E. Mazur, *Principles & Practice of Physics*, 1st ed. (Pearson, London, UK, 2014).

[52] B. D. Mikula and A. F. Heckler, Framework and implementation for improving physics essential skills via computer-based practice: Vector math, Phys. Rev. Phys. Educ. Res. **13**, 010122 (2017).

[53] M. Goos, P. Galbraith, and P. Renshaw, Socially mediated meracognition: Creating collaborative zones of proximal development in small group problem solving, Educ. Stud. Math. **49**, 193 (2002).