Establishing a relationship between student cognitive reflection skills and performance on physics questions that elicit strong intuitive responses

Cody R. Gette and Mila Kryjevskaja

*Department of Physics, North Dakota State University,
1211 Albrecht Boulevard, Fargo, North Dakota 58105, USA*

(Received 24 July 2018; published 19 March 2019)

After targeted instruction designed to improve student conceptual understanding of physics, a significant fraction of students are not able to answer many questions in a consistent manner. Prior research suggests that even those students who demonstrate that they acquired the relevant knowledge and skills (i.e., possess the requisite “mindware”) still tend to rely on their intuitively appealing (and often incorrect) ideas. This study aims to provide insights into cognitive mechanisms that may lead to the identified inconsistencies in student reasoning. We present results of an empirical investigation guided by dual process theories of reasoning and accompanying theoretical constructs of cognitive reflection and mindware. Specifically, we proposed a set of hypotheses to establish a link between student abilities to mediate intuitive responses and performance on physics questions that elicit strong intuitive responses. The cognitive reflection test (CRT), developed by cognitive psychologists, was used to measure students’ ability to engage in analytic processing in a more productive manner (i.e., cognitive reflection skills). Additionally, we developed a set of screening-target questions in the context of Newton’s third law to assess student reasoning approaches in physics. Results suggest that, in the presence of the necessary mindware, those students who possess a higher level of cognitive reflection skills are more likely to (i) arrive at a correct answer on a question that tends to elicit a strong intuitive, but incorrect response; (ii) recognize the need for justifying their answers, even if a correct answer does not require rejecting an intuitively appealing response; and (iii) engage in consistent reasoning.

DOI: [10.1103/PhysRevPhysEducRes.15.010118](https://doi.org/10.1103/PhysRevPhysEducRes.15.010118)

I. INTRODUCTION

Novice physics learners often reveal naive ideas and notions about how “things work” in physics. Some of these ideas are intuitively appealing and persist even after targeted instruction. For example, students in introductory mechanics courses often predict that “during a collision, a heavier object exerts a greater force,” and the notion that “the current is used up” is familiar to experienced instructors teaching electric circuits. Many student ideas of this nature are well documented in the physics education research (PER) literature and appear in nearly every topic across introductory physics (and beyond) [1–3]. However, recent studies revealed that multiple refinements of instructional materials and strategies designed to improve student understanding of relevant concepts do not necessarily lead to significant improvements in student performance on tasks that elicit strong intuitive responses. These studies

also indicate that a fraction of the students who demonstrate that they acquired correct formal knowledge and skills as a result of instruction, still tend to rely on incorrect (but intuitively appealing) thoughts [4–8]. In some cases, these students do not find the need to apply formal knowledge to justify their predictions. In other cases, they tend to employ incorrect reasoning to support an answer that they already believe to be correct. The latter is consistent with the “answer first, reasoning next” thinking pattern [6].

The study presented here aims to provide insights into cognitive mechanisms that may lead to inconsistencies in student reasoning. We present results of an empirical investigation guided by dual process theories of reasoning (DPTOR) and accompanying theoretical constructs of cognitive reflection and “mindware.” Specifically, DPTOR, developed and widely used in psychology research, suggest that human cognition relies on two distinct processes: Process 1 (often referred to as the *heuristic process*) and process 2 (the *analytic process*) [9,10]. The heuristic process is fast, automatic, and subconscious. It is responsible for generating a first available mental model of a situation at hand based on the reasoner’s prior knowledge, contextual cues, and overall plausibility. The analytic process, on the other hand, is slow, deliberate,

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

and effortful. In everyday life, first available responses generated by the heuristic process are often described as “first impressions” or “gut feelings.” In this study, we refer to such responses as a reasoner’s intuition.

Three aspects of DPTOR are particularly relevant to the current investigation. First, reasoners see the world around them through the lens of the fast and intuitive heuristic process. Only after a first available response is generated might the analytic process intervene and provide a thorough evaluation of the validity of that response. As such, according to DPTOR, a reasoner’s intuition cannot be turned off. Second, if a reasoner does not perceive any threat to the validity of the first available response, the analytic process is circumvented entirely. This direct reasoning path from the first available response to the final inference is often described as “cognitive miserliness” [11–13]. Finally, even when the analytic process is engaged, it operates based on biases of its own. For example, reasoners often exhibit confirmation bias and will build an argument that supports an intuitively appealing response instead of checking for consistency between this response and relevant formal knowledge, or looking for alternative solutions [14,15].

The mechanistic reasoning paths outlined by DPTOR (and further discussed in Sec. II) suggest that reasoners who are able to identify their own intuitive thoughts and engage in an effortful, logic-based analytical evaluation of such thoughts are more likely to reason productively. The ability to mediate intuitive thinking by reasoning more analytically is often called *cognitive reflection*. The cognitive reflection test (CRT), developed by Frederick, has been used to gauge this ability [16]. A key feature of the test is that for many reasoners each question immediately elicits an intuitively appealing, but incorrect, response. In order to answer correctly, a reasoner must be able to recognize the response as intuition driven, check the response for validity, and replace it with a correct solution. Generally, this solution is obtained without gaining any further knowledge. The following CRT question could be used to illustrate this reasoning path: “A bat and a ball cost \$1.10 in total. The bat costs \$1 more than the ball. How much does the ball cost?” For most people, an intuitive answer of “10 cents” springs to mind; in fact, many quickly embrace this answer as correct without further consideration [10,16]. Upon reflection, however, one may realize (without necessarily gaining additional insight into the relevant arithmetic) that the correct answer is “5 cents.”

The example above also highlights that, for some types of questions, an incorrect response does not necessarily reflect the lack of relevant knowledge and skills, but is rather indicative of reasoning tendencies. Indeed, consistent with DPTOR is an assertion that learners may reason incorrectly because (i) they lack sufficient knowledge of the rules and concepts needed to reason through a given problem and/or (ii) they behave as “cognitive misers.” In

an analogy with computer software, Stanovich refers to the “knowledge of rules and concepts” as mindware [17].

While possessing mindware is necessary, it may not be sufficient to answer questions of certain types correctly. Since the first available intuitive response (generated by the heuristic process) appears to be an entry point into a reasoning path, we hypothesized that students who possess a higher level of cognitive reflection skills are more likely to be successful at answering certain types of physics questions correctly and consistently. We argue that taking into account the interplay between intuitive and more deliberate processing, combined with cognitive reflection, may be critical for understanding and addressing persistent incorrect responses as well as inconsistencies in student reasoning on many physics tasks.

In order to ensure that a specific reasoning approach employed by a student is not due to a lack of relevant formal knowledge and skills, but is rather indicative of reasoning tendencies, we used a “screening-target questions” methodology. We chose Newton’s third law as the context for this investigation due to its ubiquity in physics instruction and a large body of PER literature documenting student thinking in this context. The screening question was designed to probe whether or not a student possesses the relevant mindware and is able to apply Newton’s third law correctly in a challenging situation that does not necessarily elicit a strong intuitive response. Tasks known to elicit intuitive responses in this context were used to inform the design of “target” questions. Two types of target questions were developed, both involving collisions between objects of unequal masses. Target question 1 elicits an intuitively appealing, but incorrect, response. Target question 2, on the other hand, elicits an intuitively appealing response that happens to be consistent with the formal analysis of the presented situation. We then analyzed, in detail, responses to the target questions from those students who answered the screening question correctly, thereby demonstrating the presence of the relevant knowledge and skills (i.e., the requisite mindware). Three hypotheses were proposed and tested.

Hypothesis 1.—Given that students answer a screening question correctly, students who possess higher cognitive reflection skills (as measured by the CRT) are more likely to provide correct answers to target question 1, which tends to elicit an intuitively appealing (but incorrect) response that must be rejected.

Hypothesis 2.—Given that students answer a screening question correctly, students who possess higher cognitive reflection skills (as measured by the CRT) are more likely to recognize the need for justifying their answers to target question 2, even if the correct answer does not require a rejection of the first available intuitive response.

Hypothesis 3.—Students who possess higher cognitive reflection skills (as measured by the CRT) are more likely to answer the screening-target questions consistently

(i.e., provide correct answers to all questions and include correct reasoning on target questions).

Section II of this paper provides further discussion of the relevant theoretical background, including interactions between the heuristic and analytic processes, cognitive reflection, and mindware. Section III includes description of our research methodologies, data collection, and analysis. Our findings and their interpretation are included in Secs. IV and V.

II. THEORETICAL FRAMEWORK AND PRIOR RESEARCH

One representation of the interaction between the fast heuristic process and the slow analytic process is illustrated in Fig. 1. When confronted with a new situation, the heuristic process quickly and subconsciously generates a mental model based on contextual cues and past experiences [9]. We refer to this initial mental model as an intuitive response. (This entry point into a reasoning path is illustrated by the upper left box in Fig 1.) In some cases, the analytic process evaluates the intuitive response, while in others, it does not. If the reasoner perceives the intuitive response to be highly plausible, the analytic process is entirely circumvented and the first available mental model yields a final answer. The term cognitive miserliness has been coined in psychology to describe this reasoning shortcut from the first available mental model directly to the final inference [13,18,19]. The analytic process could be engaged either by imposed external conditions (e.g., requirements to provide an explanation) or in the presence of a strong “red flag.” However, even if engaged, the analytic process is subject to biases of its own. It is generally “lazy,” and does not always employ well-defined criteria for accepting or rejecting mental models. Instead of systematically analyzing a presented situation, reasoners tend to default toward accepting the initial model as valid by justifying what they already believe to be true (confirmation bias) [14,20]. Only in the presence of a strong red

flag may the analytic process become fully engaged and may faults in reasoning be detected. If the analytic process does reject a response, the heuristic process suggests a new, modified mental model, and the cycle repeats, as shown in Fig. 1.

The mechanistic reasoning paths outlined by dual process theories highlight a critical aspect of human cognition: intuition cannot be turned off. Reasoners’ quick and subconscious ideas serve as an entry point into a reasoning path. As such, it is critical to direct our efforts toward understanding the factors that may mediate intuitive thinking by engaging the analytic process in a more productive manner.

In order to gauge a reasoner’s tendency to engage the analytic process and evaluate the output of the intuitive process, cognitive psychologists developed the cognitive reflection test (CRT), presented in Fig. 2 [16]. As discussed in the introduction, the key feature of each CRT question is that, for most reasoners, an intuitively appealing but incorrect answer comes to mind immediately and often without conscious effort. This answer must first be rejected and then replaced with a correct solution. (The correct answers to the three CRT questions are 5 cents, 5 minutes, and 47 days, respectively.)

The CRT has been validated and widely used in psychology research. Researchers in cognitive psychology have conducted numerous studies in order to pinpoint the specific aspects of cognition measured by the CRT. It is generally agreed that the CRT is *not* a measure of cognitive abilities or mathematical skills [17,21–23]. Toplak *et al.* probed the relationship between the CRT and performance on a wide array of tasks from the heuristics-and-biases literature selected to “reflect important aspects of rational thought,” including probabilistic reasoning, hypothetical thought, theory justification, scientific reasoning, and the tendency to think statistically [13]. Results indicate that the CRT is a more “potent predictor” of performance on heuristics-and-biases tasks than measures of cognitive ability, other thinking dispositions, or executive functioning. The researchers argued that these measures do not gauge the tendency “toward miserly processing in the way that the CRT does.” A score of 0 or 1 on the CRT indicates a disposition toward miserly processing, while a score of 2 or 3 is more indicative of a tendency to mediate intuition with conscious reflection.

FIG. 1. Interactions between the heuristic and analytic processes [9].

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. 2. Cognitive reflection test.

A growing body of PER literature provides compelling evidence suggesting that successful reasoning on some types of physics questions, much like a strong CRT performance, requires students to override initial intuitive responses. As such, it is reasonable to predict a positive relationship between CRT scores and student performance on these types of questions. At the same time, the ability to recognize shortcuts in one's own reasoning *alone* is not enough to ensure strong performance in physics. In the presence of a “mindware gap,” even if a student recognizes the intuitive nature of a compelling response, the lack of relevant knowledge is likely to prevent the productive engagement of the analytic process, thus, yielding a poor response. Our prior study, aimed at probing the relationships between cognitive reflection and student performance, provides evidence for this claim in the context of introductory physics instruction. Specifically, we found that a correlation exists between students' abilities to mediate intuitive thinking, as measured by the CRT, and student performance in physics, as measured by the Force and Motion Conceptual Evaluation (FMCE) [24]. This finding is consistent with the results by Wood *et al.*, who also found a similar relationship between CRT scores and performance on the Force Concept Inventory (FCI) [25]. Perhaps more importantly, however, our analyses also revealed that the correlation between student performance on the FMCE and cognitive reflection skills becomes significantly stronger as student understanding improves over the course of the semester. The latter result is entirely consistent with Stanovich's construct of the mindware gap [17,26].

In the present study, we build on prior research to probe the role of cognitive reflection in student reasoning in physics even further. Specifically, we aim to investigate whether or not, in the presence of relevant mindware, differences in student reasoning approaches could be accounted for by the tendency toward cognitive reflection. We are particularly interested in probing student reasoning on tasks that elicit strong intuitive responses. Student difficulties with the application of Newton's third law in the context of collisions between two objects of different masses (and/or sizes) are well documented. Therefore, student reasoning with Newton's third law was chosen as a context for the current investigation.

Work dating back to the early 1980s by Halloun and Hestenes on “common sense concepts” and by Brown on “student preconceptions” documented many student ideas regarding collisions and Newton's third law [27–29]. Hestenes described Newton's third law as “inconsistent with common sense intuitions.” He observed that most students adhered to some sort of *dominance principle* according to which a greater mass exerts the greater force or the object causing the motion exerts the greater force on the other object. Brown observed that students have a general view of objects “having” more or less force and

being more or less “force-full.” Students exhibited particular difficulties with collisions between a moving, massive object (e.g., bowling ball) and a stationary, less massive object (e.g., bowling pin) with only $\sim 5\%$ of students successfully answering after fairly traditional instruction. Maloney investigated “rule-governed approaches” to collisions (and other scenarios) involving applications of Newton's third law [30]. He reported that novice students most commonly argued that (i) mass is the determiner, meaning greater mass produces greater force; (ii) for systems at rest, all forces are equal, but for moving systems the greater mass means greater force; and (iii) for systems at rest, all forces are equal, but for moving systems the ‘cause’ exerts a greater force. Experienced students, those who previously completed a course in college physics, tended to argue that (iv) for systems at rest or moving at a constant velocity, the forces are equal; for accelerating systems, greater mass exerts greater force; and (v) for systems at rest or moving at a constant velocity, the forces are equal; for accelerating systems, the cause exerts a greater force.

The FMCE and the FCI are two concept inventories most commonly used to assess the effectiveness of instruction in introductory mechanics courses. Not surprisingly, both concept inventories include questions involving collisions [31,32]. Researchers in PER also utilize these scenarios to probe various aspects of teaching and student learning. Hammer and Elby used epistemological resources as a framework for gaining insights into student thinking, which subsequently led to the development of instructional materials designed to help students refine their everyday intuition in the context of collisions and Newton's third law [33,34]. A study by Maries and Singh focused on probing the pedagogical content knowledge of teaching assistants in an introductory mechanics course. The researchers found that some teaching assistants tend to perceive the more massive or more “active” object as exerting a greater force [35]. They noted that “The fact that even some experts hold this alternate conception after many years of practicing physics points out how strong this alternate conception is and how difficult it is to overcome it in this particular context.” Researchers in PER often interpret these responses through other theoretical frameworks such as “framing” or “resources.” We argue that DPTOR can encompass these ideas. Through the lens of DPTOR, the subconscious and automatic selection of a “frame” or a particular “resource” occurs through the heuristic process (see also Refs. [4,36]).

III. RESEARCH METHODOLOGY

In this section, we discuss in detail the “screening-target questions” methodology and describe how prior research informed the development of two specific versions of the collision task that served as target questions in this investigation. Additionally, we discuss the instructional

context for this investigation and describe data collection approaches.

A. Instructional context

This study was conducted in two semesters of a large-enrollment introductory calculus-based mechanics course taught by different instructors. In both semesters, the screening-target questions, discussed in detail below, were administered on a single exam upon completion of all relevant instruction on forces and Newton's laws. Both instructors used research-based, active learning strategies and employed two undergraduate learning assistants. The instructors implemented materials from the *Tutorials in Introductory Physics* developed by the University of Washington Physics Education Group (UW PEG) in an interactive lecture format [37,38]. Scenarios involving two colliding objects of different masses are not included explicitly into the sequence of UW tutorials on forces. However, both instructors incorporated a multiple-choice "clicker" question featuring a heavy truck colliding with a small car. Students considered multiple scenarios (e.g., before the collision, the truck moved toward the car and the car was stationary; before the collision, both vehicles were moving). They were asked to choose conditions under which the magnitude of the force that the truck exerted on the car was *greater* than the magnitude of the force that the car exerted on the truck. Students were expected to recognize that, according to Newton's third law, the two forces must be equal in magnitudes under all possible circumstances.

We note that during instruction, students were encouraged to share their reasoning to conceptual questions discussed in class. Students were asked to provide all the steps necessary to convince their peers or an instructional team member that the argument is correct and complete. In those cases when students' arguments were based on gut feelings, they were encouraged to consider how their responses were (or were not) consistent with the formal knowledge acquired in class. This instructional approach aimed to support student intuition by seeking consistency between intuitive ideas and formal physics

knowledge. While no explicit prompts were included on the course exams requesting that students provide answers based on formal knowledge (e.g., Newton's laws), it was articulated that, in order to receive full credit on free-response questions, the students were expected to apply formal reasoning approaches practiced in class.

B. Overview of the screening-target questions

This investigation aimed to probe whether or not, in the presence of relevant mindware, differences in student reasoning approaches could be accounted for by the tendency toward cognitive reflection. In order to disentangle (to the extent possible) student knowledge and accompanying skills from reasoning tendencies, we applied the screening-target questions methodology.

Screening question.—The screening question, shown in Fig. 3, requires students to recognize the applicability of Newton's third law in a situation that is challenging but not associated with a strong intuitive response. Students were asked to consider two identical blocks on a plate lifted upward with the acceleration of $g/2$. Students compared the magnitude of the normal force exerted on block A by block B (N_{AB}) to the magnitude of the normal force exerted on block B by block A (N_{BA}). In order to answer correctly, students must recognize that the two normal forces constitute a Newton's third law force pair and, therefore, their magnitudes must be equal.

Prior research on a similar question conducted by Stetzer and UW PEG suggests that, even though this context does not elicit strong intuitive responses of the same nature as those observed with two colliding objects, this screening question could still be challenging for many students [39]. In particular, after traditional lecture instruction (but before tutorial instruction), most students do not seem to have significant difficulties with the application of Newton's third law in cases when the system is either at rest or moving vertically with a constant speed ($\sim 95\%$ and $\sim 70\%$ correct, respectively). However, when presented with the system accelerated upward, only about 25% of students successfully recognize the applicability of Newton's third law. Perhaps distracted by irrelevant information

Consider the system shown at right. Blocks A and B each have mass m . Plate C, all cables and supports are massless.

The system is lifted with acceleration $g/2$ upward. Ignore air resistance.

Let N_{AB} be the magnitude of the normal force on A by B. Let N_{BA} be the magnitude of the normal force on B by A. Which of the following best represents the comparison of the magnitudes of these two forces?

- A. $N_{AB} > N_{BA}$
- B. $N_{AB} < N_{BA}$
- C. $N_{AB} = N_{BA}$
- D. Not enough information to determine.

FIG. 3. "Screening" question to probe whether or not a student is able to apply Newton's third law in a situation that does not elicit a strong intuitive response.

Two identical carts are on a straight, frictionless track. One of the carts (Cart A) is empty; the other (Cart B) is full of heavy rocks. Cart A is initially at rest, and Cart B moves along the track toward Cart A with constant velocity v_0 . Eventually, Cart B collides with Cart A.

During the collision, is the magnitude of the force exerted on Cart A by Cart B *greater than, less than, or equal* to the force on Cart B by Cart A? Explain.

FIG. 4. Target question 1: The mine cart question.

(e.g., “acceleration is upward”), many students become entangled in more complex lines of reasoning, which typically involve incorrect arguments similar to the following: “*Since the acceleration is in the upward direction, according to Newton’s second law, the vertical force in the upward direction must be greater in magnitude than that in the downward direction; as such, the normal force on block A is greater than the normal force on block B.*” It appears that the students who engage in this line of reasoning treat the Newton’s third law force pair as acting on the same object, which leads to the incorrect conclusion. We argue that a correct student response to the screening question may serve as an indication that the student (i) possesses knowledge of Newton’s third law and (ii) is able to recognize its applicability in a challenging situation.

Target questions.—Two target questions were designed to probe the impact of cognitive reflection skills on student reasoning approaches. Similar to the screening question, the target questions require the application of Newton’s third law. However, both target questions feature two canonical characteristics of a collision context known to elicit intuitive responses for some students: (i) one object is more massive than the other, and (ii) one object is “active” while the other is initially at rest.

Target question 1, the mine cart question, is shown in Fig. 4. The question involves a moving heavy cart colliding with a light cart initially at rest. Students are asked to compare the magnitudes of the forces that the two carts exert on each other during the collision. For some students, this question elicits a well documented intuitive idea that “during a collision, the force exerted by a heavy active object is greater than that by a light inactive object.” In order to answer correctly, these students must first reject or suppress this intuitive response, and then replace it with a formal argument based on Newton’s third law. A student response representing correct, complete reasoning would be similar to the following: “*Equal to. Since the forces*

acting on each cart are a force pair, according to Newton’s third law the forces are equal.” We also considered short descriptions such as “*Equal to. Third law pair.*” to be correct and complete. This question allowed us to test hypothesis 1.

Target question 2, the spaceship question, is shown in Fig. 5. This question also involves a collision between a massive object (a spaceship) and a small moving object (an asteroid). The spaceship is initially at rest, while the asteroid moves with a high speed. The question asks students to compare the magnitudes of the *accelerations* of the two objects during the collision. Formal reasoning requires two steps: (i) according to Newton’s third law, the forces that the two objects exert on each other during the collision are equal in magnitude; and (ii) according to Newton’s second law, the acceleration of a light asteroid must be greater than that of the heavy spaceship. A student response representing correct and complete reasoning would be similar to the following: “*the spaceship is less than the asteroid because the force between the two is the same. Since the asteroid has less mass than the spaceship, the acceleration will be greater.*” As in target question 1, the application of Newton’s third law is required to answer this question correctly. (While target question 2 requires one extra step, we did not anticipate significant difficulties with a straightforward application of Newton’s second law in this context.) In contrast to target question 1, the correct formal response in this scenario is also consistent with an intuitively appealing answer. Specifically, intuitively, it does not seem likely that the motion of the heavy spaceship would be affected by the collision with a light asteroid in a significant way. Instead, it is appealing to conclude that the acceleration of the heavy spaceship will be less than that of the asteroid. For most students, the path to the correct answer does not require the rejection of a first available intuitive response. We hypothesized, however, that those students who possess high cognitive reflection skills would

A spaceship is drifting through outer space with its engine off. Suddenly, a small asteroid smashes into the side of the spaceship. (Luckily, the asteroid does not smash *through* the spaceship.)

During the collision, is the magnitude of the acceleration of the spaceship *greater than, less than, or equal* to that of the asteroid? **Explain.**

FIG. 5. Target question 2: The spaceship question.

be more likely to provide a correct and complete justification for their answers. We argued that these students may have developed the habit of mind to check for consistency between formal reasoning and intuitively appealing responses (even if these responses appear to be highly plausible). As such, in our analysis, we distinguished between those responses that were correct with correct and complete reasoning, and those that were correct with incomplete reasoning. The responses in the latter category typically did not include any reference to Newton's third law, as illustrated by the following two examples. One student wrote *"Less than, because the asteroid has less mass than the spaceship. The asteroid is slowed down to the same velocity (or negative velocity) while the spaceship increases by a much smaller velocity."* Another student argued *"Less than. The asteroid has less mass and $a = F/m$."* In these responses, no reference was made to Newton's third law, or equal forces between the two objects during the collision. In the first example, the student did not appear to apply formal reasoning to the problem and instead reasoned with "everyday language" while in the second, the student appears to make an assumption that the forces were equal, but did not provide any justification for this assumption. Target question 2 allowed us to test hypothesis 2.

Student responses to all screening-target questions were used to test hypothesis 3. We consider responses to be consistent if they include correct answers to all three Newton's third law questions and correct justifications for those answers on the two target questions.

All the questions used in this study were reviewed and validated by experts (learning assistants, research team members, and the course instructors) for clarity and intended use. To ensure that students interpret the questions as intended, we reviewed student responses from prior semesters in which these tasks were used as instructional and assessment tools.

C. Data collection

The screening-target questions were included in a single exam. In order to minimize possible biases in student responses (e.g., cuing, priming), these questions were not offered in a sequence (i.e., one after another) [4,20,40–42]. Instead, they were distributed throughout the exam. Only about a third of all exam questions were offered in the free-response format, including both target questions. To avoid the placement of the three screening-target questions in close proximity to each other, the screening question was included in a multiple-choice section of the exam. This decision was supported by results of prior research suggesting that correct answers to the screening question are typically accompanied by correct reasoning. At this point in our investigation, it is not clear whether or not the results would be altered if the three research tasks were offered one after another. However, by implementing the current

research design we attempted to ensure that an entry point into a reasoning path on each question is not immediately influenced by a response on a preceding, analogous task.

In addition to the graded work (e.g., exams and homework), students were asked to participate in regular ungraded pre-tests administered in a web-based format outside the class. Pre-tests served various purposes with the overarching goal of providing formative assessment to both the instructional team and the students in the course. As such, the students were accustomed to being engaged in a range of activities that probed their understanding of various aspects related to the course. For example, some tasks required students to interpret graphs or apply specific mathematical techniques. The CRT was included on one of the regular pretests. No special instructions were provided for the students. The students were informed that the pretest continues on the next page where a set of CRT questions was included. We have no reason to believe that the students might have perceived the CRT questions as being outside of the range of typical pretest activities and, therefore, became alarmed by the inclusion of (what may appear to be) nonphysics specific questions. If the students sensed an unusual nature of the CRT questions, they may have been more likely to answer the questions correctly since the analytic system operates more productively in the state of cognitive strain (e.g., being alarmed or confused). This could have diminished the ability of the CRT to discriminate reasoners' cognitive abilities. In that case, a correlation between cognitive reflection and performance on physics problems would be more difficult to detect. We note that even if a subset of students had been exposed to the CRT questions prior to this study, recent research suggests that there is no significant negative impact on the test's predictive validity even after multiple exposures to the CRT questions [43].

IV. DATA ANALYSIS AND RESULTS

Next we discuss techniques used for testing the three hypotheses proposed in the study. First, we established criteria for categorizing student responses to the target questions, as discussed below. A subset of questions was then analyzed by two researchers to ensure reliability. Second, we established the link between CRT scores and student reasoning approaches by comparing the CRT scores of those students who succeeded on each task and those who did not. Finally, we conducted logistic regression analysis in order to determine the relationship between the likelihood (or probabilities) of successful reasoning and CRT scores.

A. Establishing the link between CRT scores and student reasoning approaches

The average CRT score for the student population in this study is $\langle \text{CRT} \rangle = 1.8$. This number is lower than that

TABLE I. Average CRT scores. For target question 2, Correct w/IR represents correct with incomplete or incorrect reasoning (IR) and Correct w/CR represents correct with correct reasoning (CR).

	Incorrect	Correct w/CR	U	p	CL
Target question 1 ($N = 122$)	$\langle 1.50 \rangle$	$\langle 1.96 \rangle$	1714	0.038	0.62
	Correct w/IR	Correct w/CR	U	p	CL
Target question 2 ($N = 98$)	$\langle 1.64 \rangle$	$\langle 2.09 \rangle$	1291	0.044	0.62
	Inconsistent	Consistent	U	p	CL
All tasks ($N = 122$)	$\langle 1.54 \rangle$	$\langle 2.13 \rangle$	2425	0.002	0.65

reported by Wood *et al.* ($\langle \text{CRT}_{\text{Wood}} \rangle = 2.4$ and somewhat higher than the average CRT score of a general population measured by Frederick ($\langle \text{CRT}_{\text{Frederick}} \rangle = 1.24$) [16,25].

In both semesters, 144 students provided responses to all research tasks used in this study (i.e., the screening-target questions and the CRT). Approximately 85% of these students ($N = 122$) answered the screening question correctly. As stated above, since our goal was to probe the link between cognitive reflection skills and student reasoning approaches in the presence of relevant mindware, in the following analysis we only included responses from those students who answered the screening question correctly. Specifically, of those students, only $\sim 75\%$ answered target question 1 correctly with correct reasoning. This result indicates that, although the majority of students were able to recognize the applicability of Newton's third law to a challenging screening question, approximately a quarter of these students failed to do the same on a question that elicited a strong intuitive response (e.g., “ N_{BA} is greater than N_{AB} because cart A has much less mass than cart B.”).

The average CRT score of the students who answered target question 1 correctly ($\langle \text{CRT}_{\text{Correct}}^{\text{Target 1}} \rangle = 1.96$) is higher than that of those students who answered incorrectly ($\langle \text{CRT}_{\text{Incorrect}}^{\text{Target 1}} \rangle = 1.50$). Statistical analysis revealed that the difference is significant with a common language effect size (CL) of 0.62 (Mann-Whitney $U = 1714$; two-sided $p = 0.038$; see Table I). According to McGraw and Wong, the common language effect size is the probability that a case randomly selected from one group will have a higher score than a case randomly selected from the other group [44]. A CL of 0.62 is comparable to a Cohen's d of 0.4 and thus can be considered a “medium” effect [45,46]. These results provide evidence in support for hypothesis 1 that, in the absence of a mindware gap, a link exists between the student success on target question 1 and cognitive reflection skills.

Approximately 80% ($N = 98$) of the students who answered the screening question correctly provided correct answers to target question 2, the spaceship task. However, only $\sim 70\%$ of these students included a correct and complete argument in support of their answer. Statistical analysis revealed a significant difference in the distribution of CRT scores of those students who perceived the need to justify

their answers and those who did not, as shown in Table I ($U = 1291$; two-sided $p = 0.044$; CL = 0.62). This result provides evidence for hypothesis 2 suggesting that reasoners who tend to engage in analytic thinking are more likely to analyze a given problem beyond an intuitively “obvious” answer or a response that “just feels right.”

Further analysis revealed that, of those students who answered the screening question correctly, only $\sim 50\%$ gave correct and complete responses to both target questions. The average CRT score for the students who answered all questions consistently (i.e., provided correct answers to all tasks with correct and complete reasoning on both target questions) is $\langle \text{CRT}_{\text{cons}} \rangle = 2.13$, while the average CRT score for those who provided inconsistent responses is $\langle \text{CRT}_{\text{incons}} \rangle = 1.54$ ($U = 2425$; two-sided $p = 0.002$; CL = 0.65). This result provides evidence in support of hypothesis 3.

The results above suggest a link between cognitive reflection skills and student reasoning on questions that tend to elicit strong intuitive responses. Specifically, those students who tend to provide correct, complete, and consistent arguments also tend to receive a higher CRT score. In order to probe this link even further, we performed logistic regression analysis that allowed us to compare the likelihood (or probabilities) of successful reasoning for students with higher and lower cognitive reflection skills.

B. Logistic regression analysis

Logistic regression is used when the dependent (outcome) variable is dichotomous (i.e., 0 or 1). The independent (predictor) variables can be of any kind. Logistic regression is utilized in many fields including machine learning, medical fields, and social sciences. Typically, it is used to predict the probability of occurrence of a specific event (e.g., pass or fail, correct or incorrect, survived or perished). The regression algorithm generates the coefficients ($\beta_0, \beta_1, \dots, \beta_k$) for each predictive variable x_k , which could be used to determine the logarithm of odds that a specific event will occur:

$$\log(\text{odds}) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k. \quad (1)$$

In Eq. (1), the odds are defined as

$$\text{odds} = \frac{p}{1-p} = \frac{\text{probability of an event occurring}}{\text{probability of an event not occurring}}. \quad (2)$$

Therefore, the probability for the model with a single predictor x takes the form

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}. \quad (3)$$

In this study, three logistic regression models were generated. In two models, the CRT score was used to predict the probability that each target question was answered correctly with correct reasoning. In the third model, the CRT score was used to predict the probability of consistency in student responses. As shown in Table II, the regression coefficients for the CRT score are statistically significant in all three models. However, the interpretation of these logistic regression coefficients is not as straightforward as for the more commonly used ordinary linear regression. For example, according to Eq. (1), the logistic regression model for target question 1 suggests that an increase in the CRT score by one unit results in the 0.39 increase in the logarithm of the odds of answering target question 1 correctly. It is, therefore, more illustrative to use logistic regression coefficients either to calculate the odds ratios or to generate graphs of probability functions. (Note that it could be seen from Eq. (1) that the odds ratio is independent of the intercept β_0 , and therefore we are not concerned with its statistical insignificance. The values of the intercepts are included in Table II for completeness and transparency.)

Table III contains the odds ratios that allow for a comparison between the odds for a successful completion of a task by students with a higher CRT score and the students who scored zero. For example, the odds ratio for students who scored 3 (Odds₃) and those who scored 0 (Odds₀) suggests that the students with a CRT = 3 are ~3.2 times more likely to answer target question 1 correctly, ~3.6 times more likely to recognize the need for justifying their answers on target question 2, and ~5 times more likely to answer all questions consistently. The results suggest that the likelihood of success on each task is related to the

TABLE II. Logistic regression coefficients for three models.

Model	Predictor	Coefficient
Target question 1 ($N = 122$)	Intercept	0.44
	CRT score	0.39 ^a
Target question 2 ($N = 98$)	Intercept	-0.02
	CRT score	0.42 ^a
All tasks ($N = 122$)	Intercept	-0.92 ^b
	CRT score	0.53 ^b

^a $p < 0.05$.

^b $p < 0.01$.

TABLE III. Odds ratios comparing the odds of answering each N3L question. Comparisons are made between CRT scores of 1, 2, 3, and CRT = 0.

	Odds ratio		
	Odds ₁ /Odds ₀	Odds ₂ /Odds ₀	Odds ₃ /Odds ₀
Target question 1	1.5	2.2	3.2
Target question 2	1.5	2.3	3.6
All tasks	1.7	2.9	5.0

tendency of the individual to mediate intuitive thoughts by engaging in a more productive analytic reasoning. Equivalently, this relationship can be visualized by plotting the probability [determined from the odds, $\text{odds} = p/(1-p)$] vs CRT score for each problem, as shown in Fig. 6. All three probability curves follow roughly identical trends, suggesting similar dependency on a CRT score. For example, for the two target questions, the probability of success on each question for the students with a CRT score of 3 is close to 0.8, while the probability of success for those with a CRT score of 0 is between 0.5 and 0.6. The probability for consistency in student responses is also higher for the students with a higher CRT score. The systematically lower probabilities for target question 2 and for consistency reflect the lower student success rate.

It is important to note that measures of fit and tests of significance for logistic regression are not identical to those used for ordinary linear regression. While in ordinary linear regression, R^2 is used as a single, widely accepted measure of goodness of fit of the model, no common measure of the goodness of fit exists for logistic regression. In our study, two statistical techniques were employed. First, the likelihood ratio χ^2 test for overall model fit was used to

FIG. 6. Target 1 and target 2 curves represent probabilities of answering each target question correctly vs CRT score. “All tasks” curve represents the probability of answering consistently.

compare the accuracy of the null model (not discriminating between different CRT scores) to that of the full model, which includes the CRT score as a predictor of student performance. Results indicate that, for all three cases, the accuracy of the full model is significantly higher than that of the null model (target 1: $\chi^2 = 4.04$, $p = 0.044$; Target 2: $\chi^2 = 3.95$, $p = 0.047$; all tasks: $\chi^2 = 9.27$, $p = 0.002$). Second, receiver operating characteristic (ROC) curves were generated for each model in order to estimate the effect size of cognitive reflection on student performance. Each model yielded the area under the ROC curve (AUC) between 0.6 and 0.65 (Target 1: AUC = 0.62; Target 2: AUC = 0.62; All Tasks: AUC = 0.65). The AUC is equivalent to the U statistic reported in Sec. IV A, and as such, the effect size calculated through the logistic regression analysis is consistent with the “medium effect size” obtained from the comparison of CRT averages [47].

V. DISCUSSIONS

The results of the data analysis provide support for the set of hypotheses outlined above. We investigated student reasoning approaches and performance on Newton’s third law tasks that tend to elicit strong intuitive responses. The results suggest that in the presence of relevant mindware, a positive relationship exists between cognitive reflection skills and student performance on those tasks. We used a screening-target methodology to disentangle (to the extent possible) student knowledge from their reasoning approaches. Specifically, student performance on the screening question facilitated identification of those students who possessed the necessary mindware to answer the target questions correctly. These students demonstrated that they are able to apply Newton’s third law correctly on a task that could be characterized as challenging, but not eliciting strong intuitive responses. We established a link between cognitive reflection skills and student performance on target questions 1 and 2 by comparing the average CRT scores of the students who answered the target questions correctly and consistently to those who did not. In particular, the average CRT score of the students who answered target question 1 correctly was significantly higher than that of those who did not. We observed that the students who answered incorrectly often appeared to fail to reject an intuitively appealing (but incorrect) response. We saw a similar relationship between the average CRT scores and student performance on target question 2. This suggests that the students with higher cognitive reflection skills are more likely to provide formal justification for their answers even if these answers do not require the rejection of the first available intuitive mental model. Finally, those students who answered all three tasks correctly and consistently also exhibited higher cognitive reflection skills (on average) compared to those students who did not succeed in doing so.

A more precise relationship between each CRT score and the likelihood (or probability) of success on each task was

established with three logistic regression models. Specifically, the logistic regression model for target question 1 suggests that a one point increase in the CRT score predicts an increase in the odds of a correct answer by a factor of $e^{0.39} \approx 1.5$. Alternatively, this result suggests that the probability of success on target question 1 increases with each increase in the CRT score. This finding provides support for hypothesis 1: in the presence of relevant mindware, on a question eliciting a strong intuitive (but incorrect) response, students who possess higher cognitive reflection skills are more likely to mediate intuitive responses by engaging the analytic process more productively. We observed nearly identical relationships between CRT scores and student success on target question 2, providing support for hypothesis 2: in the presence of relevant mindware, students who possess higher cognitive reflection skills are more likely to recognize the need for justifying their answers, even if the correct response does not require the rejection of the first intuitive mental model. The link between the cognitive reflection skills and consistency in student responses appears to be similar to (if not stronger than) that observed on each of the target questions. The logistic regression model predicts that a one point increase in the CRT score yields an increase in the odds of consistent responses by a factor of $e^{0.53} \approx 1.7$. This result provides evidence for hypotheses 3: students who possess higher cognitive reflection skills are more likely to answer the screening-target questions consistently.

In the hypothesis statements of this study, we stipulated that students must possess the mindware necessary to answer the target questions correctly. By doing so, we isolated one of the variables impacting student performance. This allowed for probing the relationships between reasoning approaches and CRT scores. It is important to note that if this stipulation is removed, then the relationships between student performance and their CRT scores are altered as well. Specifically, we do not find a relationship between student success rates on either target question and student CRT scores. These results are expected since, according to DPTOR, cognitive reflection skills *alone* are not likely to help students arrive at a correct answer. However, we did observe a correlation between students’ CRT scores and the probability that a correct answer to Target question 2 would include the reasoning necessary to justify that answer. This result is also not particularly surprising since the criteria imposed on this category of student responses (i.e., correct reasoning) essentially controls for mindware.

While we have shown that, in the presence of mindware, a correlation exists between performance on Newton’s third law tasks and cognitive reflection skills, the generalizability of these results to other specific contexts remains open for investigation. We speculate, however, that this relationship is not unique to this challenging context. Instead, the results from this study, along with prior findings, suggest that a

similar positive relationship may exist between cognitive reflection skills and student performance on many physics tasks that tend to elicit strong intuitive responses. As such, it is plausible that the development of cognitive reflection skills may play a critical role in general scientific thinking and reasoning.

VI. CONCLUSIONS

This study was motivated by the growing body of research suggesting that, even after targeted instruction designed to improve conceptual understanding of physics, many students fail to apply the acquired knowledge and skills in a consistent manner. These inconsistencies in student reasoning are particularly persistent on questions that elicit strong intuitive responses. The current study aimed to identify cognitive mechanisms that may be contributing to such inconsistencies. According to dual process theories of reasoning, a first available mental model (suggested by the fast and subconscious heuristic process) serves as an entry point into a reasoning path. As such, we hypothesized that, given that students acquired relevant knowledge and skills as a result of instruction, students with a higher tendency to mediate first available intuitive responses would be more likely to arrive at a correct answer by engaging the analytic process in a productive manner. The results of this investigation, combined with prior findings, provide evidence for the predicted positive relationship between the cognitive reflection skills and student reasoning in physics. The established relationship is particularly significant because it underlines the importance of designing instructional interventions that attend to both the development of content understanding and strengthening student reasoning skills. While the presence of relevant mindware is necessary for building a correct solution, the ability to systematically reflect on and analyze one's own reasoning path is equally critical for success. We argue, however, that instructional materials that give students explicit opportunities to examine their intuitive ideas are likely to be more effective if offered as a follow up to formal instruction designed to develop solid content understanding. Without the requisite mindware, students would

be less likely to recognize the intuitive nature of their responses or would lack formal knowledge necessary for the refinement of their intuitive ideas.

Finally, we speculate that instructional materials and approaches specifically designed to help students understand the interplay between the quick heuristic process and the slow analytic process may provide a twofold benefit to physics instruction by (i) facilitating further improvements in student performance and (ii) altering students' attitudes toward physics in general. It is generally known that many students perceive physics to be challenging (See, for example, Ref. [48]). While some students welcome the challenge, for others, physics instruction may generate apprehension and frustration. We speculate that this attitude may stem, at least in part, from a frequent struggle to reconcile the conflict between an answer that "feels right" and formal reasoning based on first principles. The feeling of unease induced by the clash between an appealing (but incorrect) intuitive response and formal reasoning could be illustrated by a statement from a student who gave correct answers to screening-target questions similar to those used in this study and provided the following reflection on his reasoning approaches: "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 the students may alleviate this struggle to some extent. It may help establish an instructional environment that emphasizes that careful examination (and possible rejection) of an intuitive response is a natural part of the reasoning process and *not* an indication of deficiency or failure. 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.

ACKNOWLEDGMENTS

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

-
- [1] L. C. McDermott and E. F. Redish, Resource Letter: PER-1: Physics Education Research, *Am. J. Phys.* **67**, 755 (1999).
 - [2] D. E. Meltzer and R. K. Thornton, Resource letter ALIP1: Active-learning instruction in physics, *Am. J. Phys.* **80**, 478 (2012).
 - [3] L. Hsu, E. Brewster, T. M. Foster, and K. A. Harper, Resource Letter RPS-1: Research in problem solving, *Am. J. Phys.* **72**, 1147 (2004).
 - [4] A. F. Heckler, The role of automatic, bottom-up processes: In the ubiquitous patterns of incorrect answers to science questions, *Psychology of Learning and Motivation* (Elsevier, New York, 2011), Vol. 55, pp. 227–267.
 - [5] M. Kryjevskaja, M. R. Stetzer, and P. R. L. Heron, Student difficulties measuring distances in terms of wavelength: Lack of basic skills or failure to transfer?, *Phys. Rev. ST Phys. Educ. Res.* **9**, 010106 (2013).

- [6] M. Kryjevskaja, 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).
- [7] M. Kryjevskaja, M. R. Stetzer, and T. K. Le, *Failure to Engage: Examining the Impact of Metacognitive Interventions on Persistent Intuitive Reasoning Approaches* (American Association of Physics Teachers, College Park, MD, 2015), pp. 143–146.
- [8] D. Sands, Evidence for the applicability of dual processing theory in physics problem solving, in *ICPE-EPEC 2013 Proceedings*, edited by L. Dvořák and V. Koudelková (MATFYZPRESS, Prague, Czech Republic, 2013), pp. 213–220.
- [9] J. S. B. T. Evans, The heuristic-analytic theory of reasoning: Extension and evaluation, *Psychon. Bull. Rev.* **13**, 378 (2006).
- [10] D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York, 2011), Vol. 1.
- [11] S. T. Fiske and S. E. Taylor, *Social Cognition*, 2nd (McGraw-Hill, New York, 1991), pp. 16–15.
- [12] P. N. Johnson-Laird, *How We Reason* (Oxford University Press, 2006).
- [13] M. E. Toplak, R. F. West, and K. E. Stanovich, The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks, *Memory Cognit.* **39**, 1275 (2011).
- [14] R. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises, *Rev. Gen. Psychology* **2**, 175 (1998).
- [15] D. Kahneman, A perspective on judgment and choice: Mapping bounded rationality, *Am. Psychol.* **58**, 697 (2003).
- [16] S. Frederick, Cognitive reflection and decision making, *J. Econ. Perspect.* **19**, 25 (2005).
- [17] K. E. Stanovich, *What Intelligence Tests Miss: The Psychology of Rational Thought* (Yale University Press, New Haven, CT, 2009).
- [18] E. J. Stupple, M. Gale, and C. 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* (Cognitive Science Society Austin, TX, 2013), pp. 388–393.
- [19] M. E. Toplak, R. F. West, and K. E. Stanovich, Assessing miserly information processing: An expansion of the cognitive reflection test, *J. Think. Reason.* **20**, 147 (2014).
- [20] C. K. Morewedge and D. Kahneman, Associative processes in intuitive judgment, *Trends Cognit. Sci.* **14**, 435 (2010).
- [21] 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.* **48**, 1 (2015).
- [22] G. Campitelli and P. Gerrans, Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach, *Mem. Cogn.* **42**, 434 (2014).
- [23] J. M. Liberali, V. F. Reyna, S. Furlan, L. M. Stein, and S. T. Pardo, Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment, *J. Behav. Decis. Making* **25**, 361 (2012).
- [24] M. Kryjevskaja, C. R. Gette, N. Grosz, M. R. Stetzer, and A. Boudreaux, Investigating student reasoning with the cognitive reflection test (to be published).
- [25] A. K. Wood, R. K. Galloway, and J. Hardy, Can dual processing theory explain physics students' performance on the Force Concept Inventory?, *Phys. Rev. Phys. Educ. Res.* **12**, 023101 (2016).
- [26] K. E. Stanovich, Rational and irrational thought: The thinking that IQ tests miss, *Sci. Am. Mind.* **20**, 34 (2009).
- [27] I. A. Halloun and D. Hestenes, Common sense concepts about motion, *Am. J. Phys.* **53**, 1056 (1985).
- [28] I. A. Halloun and D. Hestenes, The initial knowledge state of college physics students, *Am. J. Phys.* **53**, 1043 (1985).
- [29] D. E. Brown, Students' concept of force: The importance of understanding Newton's third law, *Phys. Educ.* **24**, 353 (1989).
- [30] D. P. Maloney, Rule-governed approaches to physics—Newton's third law, *Phys. Educ.* **19**, 37 (1984).
- [31] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, *Phys. Teach.* **30**, 141 (1992).
- [32] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the Evaluation of Active Learning Laboratory and Lecture Curricula, *Am. J. Phys.* **66**, 338 (1998).
- [33] A. Elby, Helping physics students learn how to learn, *Am. J. Phys.* **69**, S54 (2001).
- [34] D. Hammer and A. Elby, Tapping epistemological resources for learning physics, *J. Learn. Sci.* **12**, 53 (2003).
- [35] A. Maries and C. Singh, Teaching assistants' performance at identifying common introductory student difficulties in mechanics revealed by the Force Concept Inventory, *Phys. Rev. Phys. Educ. Res.* **12**, 010131 (2016).
- [36] C. R. Gette, M. Kryjevskaja, M. R. Stetzer, and P. R. Heron, Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy, *Phys. Rev. Phys. Educ. Res.* **14**, 010113 (2018).
- [37] L. C. McDermott and P. S. Shaffer (P.E.G. at the University of Washington), *Tutorials in Introductory Physics* (Prentice-Hall, Upper Saddle River, NJ, 2002).
- [38] M. Kryjevskaja, 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).
- [39] M. R. Stetzer (private communications).
- [40] A. A. diSessa, Toward an epistemology of physics, *Cognit. Instr.* **10**, 105 (1993).
- [41] E. T. Higgins, Knowledge activation: Accessibility, applicability, and salience, in *Social Psychology: Handbook of Basic Principles* (Guilford Press, New York, NY, 1996), pp. 133–168.

-
- [42] E. F. Redish, A Theoretical Framework for Physics Education Research: Modeling Student Thinking, [arXiv: physics/0411149](#).
- [43] M. Bialek and G. Pennycook, The cognitive reflection test is robust to multiple exposures, *Behav. Res. Meth. Instrum. Comput.* **50**, 1953 (2017).
- [44] K. O. McGraw and S. Wong, A common language effect size statistic, *Psychol. Bull.* **111**, 361 (1992).
- [45] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Routledge, New York, 1988).
- [46] W. P. Dunlap, A program to compute McGraw and Wongs common language effect size indicator, *Behav. Res. Meth. Instrum. Comput.* **31**, 706 (1999).
- [47] S. J. Mason and N. E. Graham, Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation, *Q. J. R. Meteorol. Soc.* **128**, 2145 (2002).
- [48] M. Sobel, Physics for the non-scientist: A middle way, *Phys. Teach.* **47**, 346 (2009).