Reinforcing mindware or supporting cognitive reflection: Testing two strategies for addressing a persistent learning challenge in the context of air resistance

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

Andrew Boudreaux
Department of Physics & Astronomy, Western Washington University, Bellingham, Washington 98229, USA

Drew J. Rosen
School of Physics and Astronomy, University of Edinburgh, Edinburgh, EH9 3JZ, United Kingdom

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

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

(Received 28 May 2024; accepted 6 August 2024; published 16 September 2024)

In this study, we have explored the effectiveness of two instructional approaches in the context of the motion of objects falling at terminal speed in the presence of air resistance. We ground these instructional approaches in dual-process theories of reasoning, which assert that human cognition relies on two thinking processes. Dual-process theories suggest multiple possible avenues by which instruction might impact student reasoning. In this paper, we compare two possible instructional approaches: one designed to reinforce the normative approach (improving the outputs of the intuitive process) and another that guides students to reflect on and analyze their initial ideas (supporting the analytic process). The results suggest that for students who have already demonstrated a minimum level of requisite knowledge, instruction that supports analysis of their likely intuitive mental model leads to greater learning benefits in the short term than instruction that focuses solely on providing practice with the normative mindware. These results have implications for the design of instructional materials and help to demonstrate how dual-process theories can be leveraged to explain the success of existing research-based materials.

DOI: 10.1103/PhysRevPhysEducRes.20.020116

I. INTRODUCTION

Many successful instructional interventions in physics rest on a foundation of research into student understanding of those topics. By drawing on the results of investigations into common student response patterns, curriculum developers are able to anticipate how students will respond to questions posed in instructional materials and devise materials that support correct and appropriate reasoning pathways while helping students identify the flaws in other approaches [1,2]. Within this framework, some curriculum developers focus on eliciting a common incorrect response, confronting students with the ways in which this response is inconsistent with physics formalism and guiding students through a reasoning process that will allow them to resolve this inconsistency ("elicit, confront, resolve") [3]. Others focus on "refining intuition," by following a similar approach while also explicitly attending to the development of students' epistemologies [4]. Regardless of their theoretical orientations, teachers and researchers recognize the value in identifying the ways in which students will attempt to answer a question and in leveraging that information in their instruction.

Not all PER-inspired instruction is equally successful, however. Many educators have had the experience of guiding students through a challenging problem using a carefully designed set of scaffolded questions leading to a particular conclusion, only to have those students fail to

*Contact author: bal23@psu.edu

arrive at the appropriate conclusion at the culmination of the reasoning chain [5]. Some researchers might refer to non-normative response patterns that arise frequently, even after high-quality instruction, as providing evidence of persistent "difficulties" [1,2]. Such patterns do not, however, necessarily indicate the presence of stable misconceptions. Indeed, the phenomenon of students responding inconsistently on two questions requiring identical content understanding and reasoning is well documented [6–8]. Emerging work in PER suggests that in many cases in which students commonly respond incorrectly to a given prompt, the errors are not due to a lack of content understanding but rather due to the nature of human reasoning itself [6–12]. Researchers have increasingly begun to draw from dual-process theories of reasoning, developed in cognitive psychology [13–16], to account for inconsistencies in student reasoning and have begun to use these theories to develop educational interventions and curricular materials [9,17–21].

According to dual-process theories, human cognition involves two distinct processes: an intuitive, rapid process and an analytic, deliberate process. When a question is encountered, the intuitive process generates a mental model suggesting a possible answer. This intuitive model is heavily influenced by aspects of the question itself and the framing with which the reasoner views the task [22]. The analytic process may or may not activate to evaluate this provisional model. If the analytical process remains inactive, the response based on the provisional mental model becomes the final answer. However, even if the analytic process engages, it can be influenced by various reasoning biases (such as confirmation bias) and may not always provide an accurate evaluation of the provisional model. Dual-process theories can thus account for inconsistencies between students' demonstrated content understanding of one question and their non-normative answer to another question.

In designing instructional materials, however, a question arises: How should inconsistencies between a student's response and their own knowledge be addressed? Multiple approaches are possible and consistent with dual-process theories. These include both cognitive approaches, which focus on strengthening the appropriate knowledge, and metacognitive approaches, which involve analyzing one's own thinking. In other words, the cognitive approach might support the intuitive process of generating a normative mental model consistent with the one an expert might produce, while the metacognitive approach provides the analytic process with support to engage in a sustained and productive analysis of the provisional model.

In this work, we attempt to determine which of these approaches—the cognitive or the metacognitive—is more productive for one specific context. Our primary research question is, "For students who have already demonstrated a minimum level of requisite knowledge for a particular question, to what extent is instruction that encourages students to reflect on and analyze their initial ideas for the question more or less successful than instruction designed to reinforce the normative approach to that question?" To explore this research question, we draw on dual-process theories to build an intervention in which students analyze the motion of a pair of objects falling at terminal speeds in the presence of air resistance. We conducted a controlled experiment to contrast the effects of two intervention strategies. The first approach, the treatment intervention, includes both cognitive and metacognitive components and seeks to support students in analyzing their likely intuitive responses. The alternate approach, the control intervention, is purely cognitive and provides repeated practice with a carefully crafted set of scaffolding questions designed to guide students through the correct reasoning approach. Although our treatment intervention could be seen as related to prior intervention approaches like elicit, confront, resolve, or refining intuitions, in this work, we draw explicitly on the dual-process framework to describe the mechanisms by which the treatment and control interventions may impact student performance. In addition, we replicated the experiment at a second university serving students who had received a different type and amount of instruction on terminal speed motion but who were otherwise similar to the students in the primary study population. Results help us to better understand for which student populations these approaches might be most successful.

II. THEORETICAL FRAMEWORK: DUAL-PROCESS THEORIES OF REASONING

In recent years, dual-process theories, which explain many domain-general reasoning phenomena [13–16], are receiving increasing attention for their ability to predict and explain student response patterns on questions in physics [5–10,12,19] and other disciplines [20,23]. Many studies demonstrate that students are able to draw upon relevant concepts to respond correctly to a "screening question" and yet fail to draw upon those same relevant concepts in response to a "target question" [7,24]. Dual-process theories posit two cognitive processes involved in reasoning and decision making: Process 1, which is fast, automatic, and intuitive; and process 2, which is slow, effortful, and analytic. When presented with a problem, process 1 will generate a provisional mental model without conscious effort, informed by factors including the reasoner's prior knowledge and beliefs as well as contextual cues associated with the problem itself. Process 2 may or may not engage in evaluating this model; whether it does so or not is mediated by factors including the individual's tendency to scrutinize their provisional models (i.e., their tendency to engage in cognitive reflection) [25] and the feeling of rightness about their own provisional model [26]. A critical aspect of

dual-process theories is that the quick and subconscious intuitive process cannot be turned off.

A. Reasoning pathways and hazards

Kryjevskaia *et al.* [12] use dual-process theories to describe a set of reasoning pathways students may follow in solving a physics problem, as well as hazards they may encounter along the way (see Fig. 1). When a task is initiated, the student's prior knowledge, experiences, and expectations, as well as features of the task at hand, together guide the subconscious creation of an initial, provisional mental model. The more familiar the reasoner is with the appropriate concepts and reasoning for the task (referred to as the necessary "mindware" in an analogy to computer software [27]), the more likely these concepts will be activated through process 1, leading to an accurate and productive initial mental model. This process of automatic and subconscious recognition is often referred to as intuition [28]. Indeed, for many experts, the correct reasoning pathways have been automatized: The intuition of experts is mostly accurate. Novice reasoners, however, may not have a wide range of relevant experiences with the topic, leading to their intuition being much less accurate in unfamiliar situations. This is illustrated in Fig. 1 as the first reasoning hazard a student may encounter, hazard A. The presence of highly salient but irrelevant features of the task (*salient distracting features*) [10] may increase the likelihood that process 1 generates an incorrect mental model. These features often hinder students' accurate recognition of a particular situation by overshadowing relevant knowledge and cuing an unproductive provisional mental model. In other words, salient distracting features may increase the relative cognitive accessibility of an incorrect model such that it comes to mind more easily than the correct model even if the correct model is cognitively available [22]. When a student's initial mental model is highly accessible relative to other models, that student is less likely to consider alternatives [11]. If the initial model is accepted without review by process 2, the reasoner follows the path of cognitive frugality, hazard B. Even if process 2 does activate to review the initial mental model, the analysis it provides may be biased or ineffective (hazard C). For instance, the reasoner may engage in confirmation bias, rationalizing the provisional model rather than searching for alternatives. Finally, if the requisite "mindware"—i.e., knowledge and skills relevant to the given situation—is not available to the reasoner, they may fail to detect and override any errors in the provisional mental model

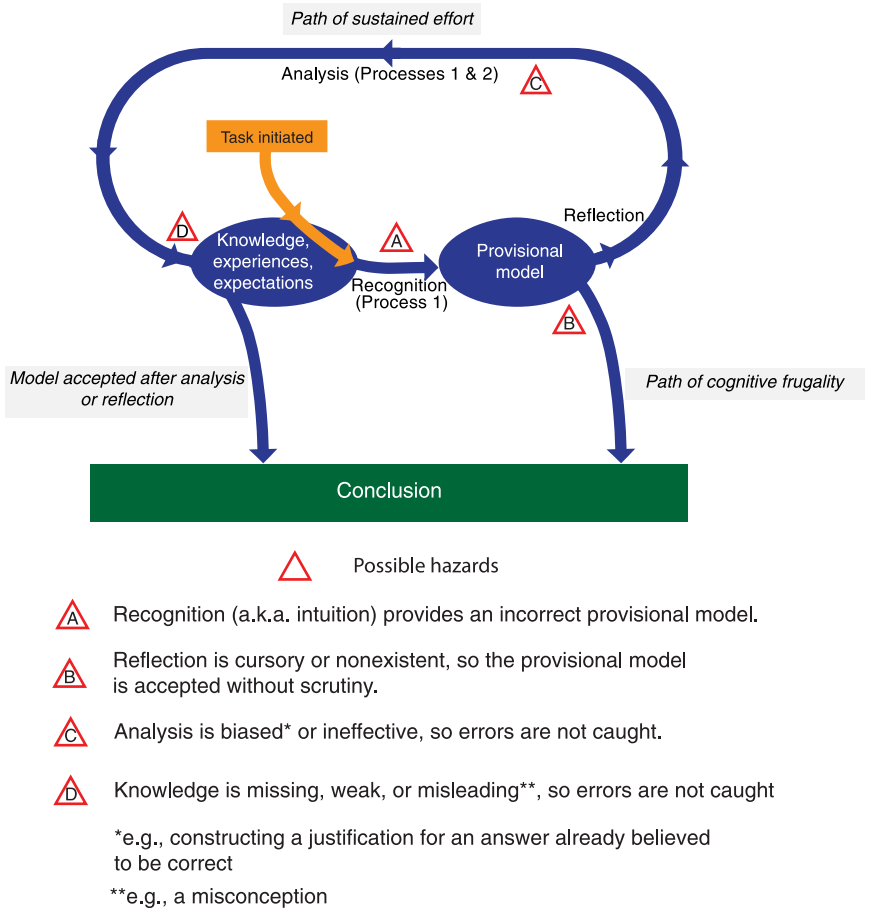


FIG. 1. Reasoning pathways and associated hazards from Kryjevskaia *et al.* [12].

(hazard D). Thus, even if the analytic process is engaged, the reasoner may accept an incorrect provisional mental model and use it to produce their final response. If the provisional model is rejected, process 1 will suggest a new model and the cycle will repeat.

B. Instructional approaches suggested by dual-process theories

A model for cognition based in dual-process theories suggests multiple possible methods for improving instruction. One approach focuses on increasing the likelihood that the intuitive process will produce a correct mental model—minimizing hazard A and thereby improving the quality of the outputs produced. Another seeks to strengthen process 2, increasing the likelihood it will engage productively to evaluate and reject incorrect mental models (minimizing hazards B and C). In this work, we present a controlled experiment testing the effectiveness of instructional strategies informed by these two different but complementary approaches.

To increase the likelihood that process 1 will produce a correct mental model, instruction might be structured to provide repeated practice with the normative mindware. This could increase the cognitive accessibility [11] of the appropriate reasoning [11,12], making it more likely that process 1 will generate a correct initial mental model—in other words, increasing the likelihood that the student will recognize the correct approach when first presented with a novel problem. In effect, this approach would be targeting reasoning hazard A in Fig. 1. As noted in Ref. [12], however, this approach would require students to be trained on a large number of similar problems and would not necessarily lead them to recognize the correct approach on a problem that involves more than minor deviations from problems seen before. Moreover, although there is evidence that training on essential skills holds the potential to increase student fluency (decreasing both the number of errors and the time it takes students to respond to certain questions) [29], many open questions remain regarding this approach, including the appropriate grain size for the knowledge on which to train students, and whether students can be trained on reasoning approaches as opposed to specific concepts or mathematical skills.

An alternative instructional strategy involves supporting a productive engagement of the analytic process. This approach targets hazards B and C in Fig. 1. Within a single context, it may involve helping students to detect an error in their provisional mental models ("raising a red flag" about a possible error) and allowing them to effectively recognize the ways in which an incorrect provisional model fails (presuming that they have access to the appropriate mindware and can thereby avoid hazard D). Since the role of the analytic process, according to dual-process theories, is to evaluate the provisional mental model, instruction that engages with students' actual provisional models (rather than just reinforcing the normative model) is a necessary component of interventions targeting hazards B and C. A "micro-intervention" approach that assists students in recognizing their errors has shown some success at improving correct response rates in the short term to a question that frequently elicits an incorrect provisional model [9]. Over time, such an approach should involve providing students with problem-solving and reasoning tools that will help them to productively analyze unfamiliar scenarios, bringing them a bit closer to "thinking like a scientist" [30].

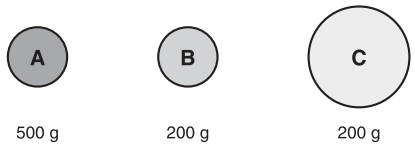
Both instructional approaches have some documented successes in different contexts [9,29]. The present study aims to compare the effectiveness of two modest instructional sequences in the same physical context. The first, referred to as the "control" strategy, targets hazard A by reinforcing the normative reasoning. The second, referred to as the "treatment" strategy, targets hazards B and C by supporting a productive analysis by process 2. Critical for this investigation was the identification of a set of research tasks in which the target question elicited predictable common incorrect intuitive ideas, and yet screening questions suggested that most students would readily demonstrate knowledge of the appropriate mindware under other circumstances (ensuring that hazard D should not be an issue). For these reasons and those documented in Secs. III and V below, the behavior of objects falling at terminal velocity due to air resistance was chosen as the instructional context.

III. EMPIRICAL BACKGROUND: RESEARCH ON STUDENT UNDERSTANDING OF TERMINAL SPEED MOTION

Results from a multiyear empirical investigation of student understanding of air resistance and terminal speed motion [31,32] served as a foundation for our development of instruction based on dual-process theories. We developed several assessment tasks to probe student understanding. In the *three-ball task*, shown in Fig. 2, students are told that balls A, B, and C are dropped from a tall building and allowed to fall. Each ball initially speeds up but eventually reaches constant speed, its terminal speed. Balls A and B have the same size but different mass, with $m_A > m_B$, while B and C have the same mass but different size. Students are asked to (a) rank A–C according to the magnitude of the air resistance force while moving at terminal speed, and (b) rank A–C according to the terminal speed.

The three-ball task has been administered in written form to students from a variety of populations (including both introductory calculus-based mechanics courses and courses for preservice teachers, at multiple institutions). In all cases, response patterns were similar: the correct response rate for the comparison of the air resistance forces was well under 50%, and explanations for both the air resistance and speed comparisons tended to be brief and incomplete.

Balls A-C are dropped from the roof of a tall building and allowed to fall vertically downward. The three balls have the sizes and masses shown in the diagram.



After being dropped, each ball initially speeds up, but eventually reaches a constant speed (its *terminal speed*). It is not known whether the balls have the same or different terminal speeds.

(a) Rank, in order from greatest to least, the magnitude of the force of air reisistance acting on each ball *while they are falling at their terminal speeds.* If any two magnitudes are the same, state that explicitly. Explain your reasoning.

(b) Rank, in order from greatest to least, the terminal speeds of the three balls. If any two are the same, state that explicitly. Explain your reasoning.

FIG. 2. The three-ball task as administered to students.

Many students did not refer to, or seem to recognize, the $F_{\text{net}} = 0$ condition when comparing the drag forces exerted on objects moving at terminal speed. Instead, students tended to compare the drag forces based on a salient feature of the objects, in this case, the relative size of the objects. The following examples are representative of student explanations for their comparison of the drag forces:

> "C has more surface area so would probably have more air resistance".
> "We have $F_{\text{drag}} = \frac{1}{2}C\rho v^2 A$, so objects that have more surface area have more drag".
> "I think the magnitude of the force of air resistance is affected by the size of the ball, not the mass".

Although a majority of students, around 75%, correctly compared the terminal speeds of balls B and C, the explanations for these responses tended to be similarly brief and incomplete (although not explicitly incorrect). While some of these explanations did mention the air resistance force, almost none referred explicitly to the specific comparison of $F_{\text{air on C}}$ and $F_{\text{air on B}}$ made in part a. The following example explanations are representative:

> "C will have the lower terminal velocity because it is larger".
> "Denser objects experience less drag, so they have less upward acceleration from F_{drag}".
> "$v_C < v_B$ because they are equal in mass, but C is larger, so it will catch more wind".

The high incorrect response rates for the comparison of the drag forces suggested to us that students were failing to apply the zero net force condition for these specific questions. As experienced instructors, however, we felt confident that our students were aware of Newton's second law and able to flexibly apply it in many situations. Dual-process theories of reasoning have provided a way to make sense of this response pattern, as will be discussed in Sec. VII.

IV. METHODS

A. Context for research

The intervention was implemented in the introductory, calculus-based mechanics course at Western Washington University (WWU). WWU is a comprehensive, regional, public, master's-granting institution that enrolls about 15 000 students and focuses primarily on undergraduate education. Introductory calculus-based physics is a three-quarter sequence that starts in the fall quarter and in the winter quarter and consists of 4 h of lecture and a 2-hr required lab. The lab materials, uniform across all sections of the course, have been developed locally over the past 15 years, based on nationally disseminated, research-based curricula including *Tutorials in Introductory Physics* [33], *Physics and Everyday Thinking* [34], and *Minnesota Context-Rich problems* [35]. The labs focus on the development of concepts and reasoning through guided exploration with relatively simple equipment and do not emphasize quantitative analysis of measurement uncertainty or the verification of physical laws.

The mechanics course covers kinematics, Newton's laws, and conservation of energy and momentum. Air resistance, including a quadratic drag force model and analysis of the terminal speed behavior of falling objects, is typically covered in a short lecture (10–20 min), along with assigned textbook reading and one or two homework problems. About 5 years ago, a 2-hr lab activity on air resistance and terminal speed was developed and included in the regular sequence of labs. Students complete this lab after most other course instruction on Newton's laws, and after the brief lecture instruction on air resistance mentioned above. A focal learning target of the lab is for students to coordinate Newton's second law and a quadratic model for the drag force to compare the terminal speeds of

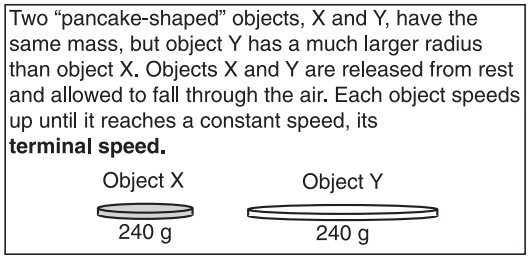


FIG. 3. The pancakes scenario. In the pancakes target question, students were asked whether the magnitude of the drag force acting on object Y is greater than, less than, or equal to the magnitude of the drag force acting on object X once each object has reached its terminal speed.

falling objects with different shapes, sizes, and masses. During lab, students examine the terminal speed behavior of a coffee filter that is released from rest and allowed to fall about 2 m to the floor. Students use an ultrasonic motion detector to generate a graph of velocity vs time and are guided to draw free-body diagrams and apply Newton's second law for an instant at which the filter is speeding up as well as an instant at which the filter is moving downward with constant speed. Students compare the behavior of two filters that have the same size and shape but different mass. In particular, students are guided to use Newton's laws and the drag force model to explain why the heavier filter has a greater terminal speed. This lab represents the primary difference between the students at WWU and the students in our replication population, described in Sec. VI.

All students in the WWU study population had completed the lab on air resistance. The intervention itself was administered as a postlab assignment. Each lab in the course includes a required, online postlab homework assignment. Credit for the air resistance postlab was entirely participation based. The postlabs are not timed or monitored. Students are told to work individually, and not consult the Internet, but are encouraged to use their notes from lab as well as other written resources from the course when completing the postlabs. The postlab assignments were due 2–3 days after students had completed the lab. We have found that student responses to open-ended questions on these assignments typically suggest that they are engaging thoughtfully with the questions as intended.

B. Design of controlled experiment

Here, we provide a broad overview of the study structure, with more details of each stage provided in the subsequent subsections. This study uses a pre- and post-test between-subjects design in which students were asked variations on the same question (the "target" question) at various points in an instructional sequence. The first variation on the target question, referred to as the "pancakes" scenario, is shown in Fig. 3. The target question required students to compare the drag forces on two objects of the same mass but different cross-sectional areas moving at terminal speed. Students were randomly assigned to one of two instructional conditions (control or treatment). The instruction in both conditions focused on helping students recognize they should apply Newton's second law to determine that the drag forces acting on two objects of equal mass falling at constant terminal speeds must be equal, as each object's drag force must equal its weight. The two conditions, however, differed in how they approached this goal as described below. The effects of the experiment were assessed after differentiated instruction using two additional variations on the target question, one (referred to as the "post-test" target question) that differed from the pancakes target question only in surface features, and another (referred to as the "near-transfer" target question) that asked students to compare the drag forces on three falling objects. The performance on these questions of students assigned to different instructional conditions was compared to determine the efficacy of the two interventions. The overall structure of the instructional and assessment sequence is shown in Fig. 4. The entire instructional package was delivered as a single online

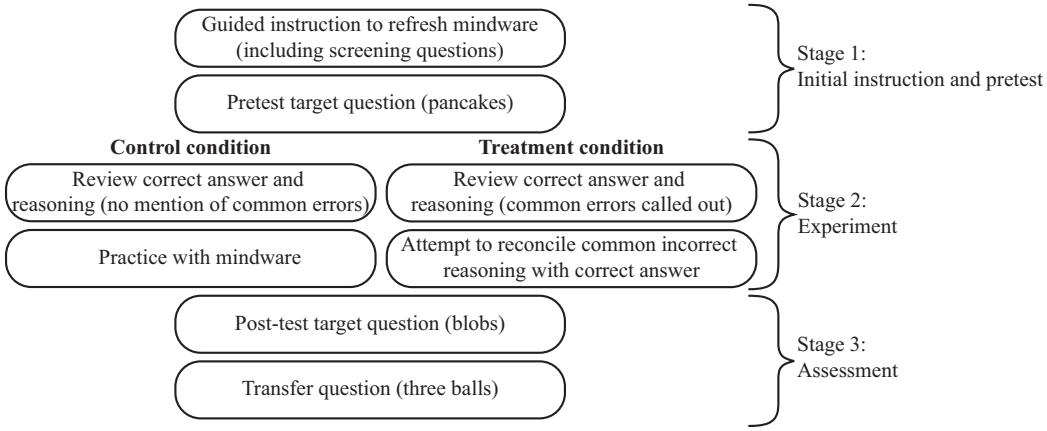


FIG. 4. Structure of the instructional and assessment sequence.

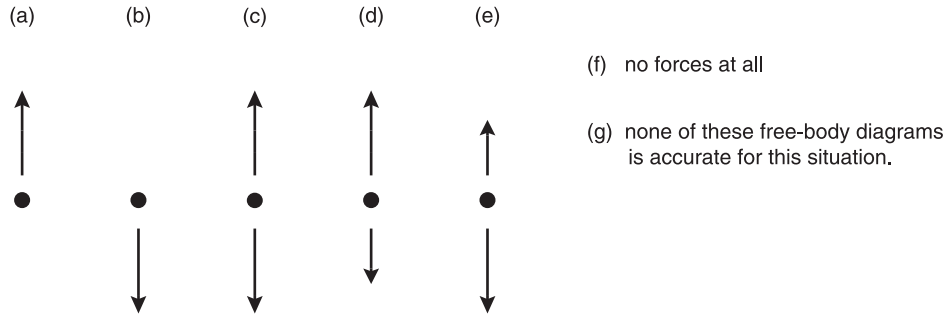


FIG. 5. Multiple-choice options from which students were expected to select the correct free-body diagram for object X when it is falling at terminal speed.

assignment via the Qualtrics platform [36]. Students were randomly assigned to either the control or treatment condition using Qualtrics' randomization functionality. The stages of the instructional and assessment sequence are described in more detail below.

1. Stage 1: Initial instruction and pretest

Stage 1 of the instructional sequence, involving the pancakes scenario (Fig. 3), was the same for students in both the treatment and control conditions. The primary emphasis of stage 1 was on a series of scaffolding questions designed to guide students to a correct comparison of the drag forces on the two objects. First, however, students were asked how the terminal speeds of the two objects would compare. We do not consider this question part of the scaffolding sequence, because most students are able to correctly compare the terminal speeds, but many do so without drawing on formal physics reasoning, as discussed in Sec. III. Instead, this question was designed to elicit the (correct) response that object X has a larger terminal speed and thereby raise the salience of the difference in the terminal speeds of the two objects to students. The first scaffolding question asked students whether the speed of object X is increasing, decreasing, or constant when it is moving at terminal speed. Students were next asked whether the acceleration of object X is zero or nonzero when it is moving at terminal speed and to explain their reasoning. They were then asked whether the net force on X is zero or nonzero and to explain. Students were next asked to select which choice in Fig. 5 best represents the free-body diagram for object X when it is moving at its terminal speed. Students were again asked to explain the reasoning for their selection. Finally, students were asked whether the free-body diagram for object Y would be different in any way from the free-body diagram for object X and to explain. Students responded to each of these questions on a single page of the online assignment and received no feedback on their responses.

To answer correctly, a student can recognize that at terminal speed, the velocity of object X is not changing. Its acceleration and the net force on it are thus zero. The free-body diagram involves two balanced forces, a gravitational (weight) force and a drag force, and the diagram for object Y would be identical to that for X. We expected these questions to remind students of how Newton's second law can be used to reason about the drag force exerted on an object moving at terminal speed. In addition to a reminder, however, a subset also served as screening questions [7,21]. As will be described in Sec. V, they were used to identify which students already could demonstrate an understanding of Newton's second law sufficient to answer the target question correctly, thus screening for availability of the appropriate mindware.

After completing the scaffolding questions, students moved to a new page on which they responded to the pancakes target question. They were again shown the image from Fig. 3 and were asked whether the magnitude of the drag force on object Y is greater than, less than, or equal to the magnitude of the drag force on object X when each object is moving at its terminal speed. On the next page, students were reminded of the answer they had given and asked to explain the reasoning for their choice.

2. Stage 2: Experiment

Stage 2 of the instructional sequence differed between the control and treatment groups. In both, students were reminded of their own answer for the target pancakes question, provided with a correct answer and reasoning, and shown the correct free-body diagrams in Fig. 6. The wording of the text describing the correct answer differed slightly between the two groups, as did the subsequent question sequences. The instruction in the control condition

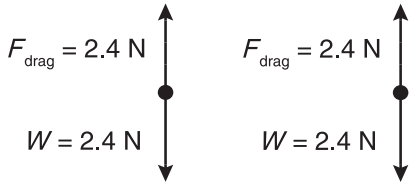


FIG. 6. Free-body diagrams for objects X and Y that were provided to students during stage 2 of the instructional sequence.

provided additional scaffolded practice with the normative mindware. This practice was intended to increase the accessibility of the mindware and target hazard A from Fig. 1. In the treatment condition, the instruction was designed to alert students to a problem with their likely incorrect response, helping them to avoid the path of cognitive frugality and avoid hazard B from Fig. 1, and support students in analyzing that response, thereby supporting process 2 and targeting hazard C. Below, we describe the stage 2 instruction for each group in detail.

Control condition. The control condition was designed to target hazard A from Fig. 1 and provide students with additional practice applying the normative physics reasoning based on Newton's second law to the case of objects falling at terminal speed. This instruction aligns with typical question sequences that might be found in instructional materials. It was expected that this additional practice would improve students' reasoning fluency and thus increase the cognitive accessibility of the correct reasoning pathway—i.e., make the normative reasoning more likely to "come to mind"—when students encounter a new situation.

In the control condition, the correct answer and reasoning for the pancakes question were simply stated with no reference to the common incorrect response. Students were then told that they would have the opportunity to practice this reasoning in a new scenario involving two falling "blobs" (see Fig. 7). Students were subsequently presented with a sequence of scaffolding questions identical to those presented previously in the pancakes scenario (see Sec. IV B 1).

Treatment condition. The treatment condition was designed to target hazards B and C and support students in the productive engagement of their analytic process by alerting students to possible errors in the common incorrect response that the object with a larger cross-sectional area would experience a larger drag force. Students were guided to reconcile their likely initial mental model with the correct reasoning for that question. The treatment sequence represented an attempt, through scaffolding, to provide students an experience of rejecting an intuitively compelling (but non-normative) model generated quickly and effortlessly by process 1 through the sustained engagement of process 2, ultimately arriving at a higher quality, normative model. It also allowed students to reflect on what the best approach for checking their provisional mental model should be.

In the treatment condition, the text describing the correct answer and reasoning for the pancakes target question was preceded by the following sentence: "Many people state that the object with the larger surface area, object Y, would experience a bigger drag force than object X because the drag force is proportional to the cross-sectional area of the object." After reading a correct explanation for the pancakes target question, students were then reminded of the drag force equation, $F_{\text{drag}} = \frac{1}{2}C\rho A v^2$, and were asked, "If it is true that $F_{\text{drag on X}} = F_{\text{drag on Y}}$, then what does the relationship between drag force, cross-sectional area, and speed ($F_{\text{drag}} \propto Av^2$) tell us about how the terminal speeds of the two objects must compare?" Students were provided with multiple-choice options from which they could indicate that the terminal speed of Y is greater than, less than, or equal to the terminal speed of X or that not enough information is available to answer. Students were then asked to explain the reasoning for their choice. This question was designed to help students realize that while the drag force equation does apply in the situation at hand, it requires students to compare both the cross-sectional area and the terminal speed of the two objects.

The treatment condition concluded with a question designed to allow students to articulate the relative merits of approaching the target questions using the drag force equation compared to Newton's second law (see Fig. 8). Students were expected to recognize that the argument based in Newton's second law must apply, whereas the approach using the drag force equation requires students to coordinate multiple variables and is thus less useful for the target question. The ultimate goal of this question was to help students be more successful in selecting and applying an approach for checking the validity of their provisional model in the future.

3. Stage 3: Assessment

The effectiveness of the instruction in the two experimental conditions was assessed using two different

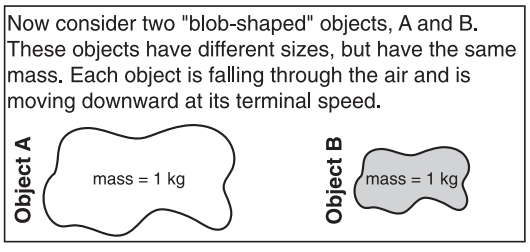


FIG. 7. The blobs scenario. In the target question, students were asked to compare the magnitudes of the drag forces on the two objects once each had reached its terminal speed.

A student is struggling with these ideas:

Student: "I think that the equation $F_{drag} = F_{gravity}$ applies. That equation leads me to conclude that $F_{drag\ on\ X} = F_{drag\ on\ Y}$ because X and Y have the same mass. I also think that the equation $F_{drag} = 1/2\ C\ \rho_{air} A v^2$ applies. That one leads me to conclude that $F_{drag\ on\ X} < F_{drag\ on\ Y}$ because Y has a bigger surface area than X. How do I reconcile these two ideas?"

What would you tell this student?

FIG. 8. A question from the treatment sequence.

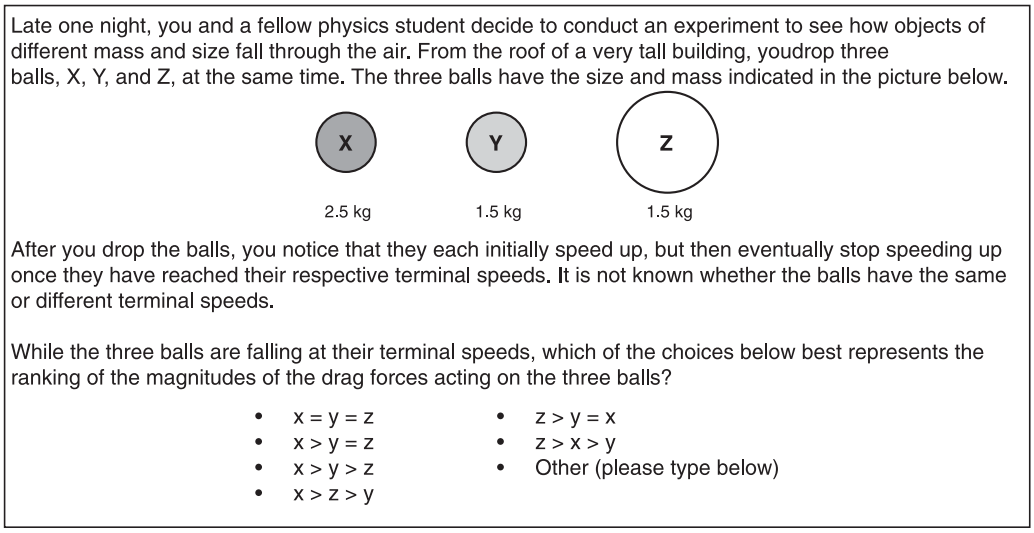


FIG. 9. The three-ball near-transfer post-test question.

questions. The blobs target question, shown in Fig. 7, was completely analogous to the pancakes target question. Students were asked to compare the magnitudes of the drag forces on the two blob-shaped falling objects after each object had reached its terminal speed. Since the two blobs have the same mass, they must also have the same drag force once they are each moving at constant speed. As noted in Sec. IV B 2, during stage 2 of the intervention, students in the control condition had answered a set of scaffolding questions for the blobs scenario similar to the ones that all students answered for the pancakes scenario in stage 1. In the treatment condition, students' first encounter with the blobs scenario was in stage 3 of the intervention.

The "three balls" post-test question (see Fig. 9) served as a near-transfer question. It required students to rank the drag forces on three falling spheres after each had reached its terminal speed. Spheres X and Y have the same cross-sectional area but different masses, while sphere Z has the same mass as Y but a larger cross-sectional area. At terminal speed, the drag force on each sphere is equal to its weight, and so the correct ranking of drag forces is identical to the ranking of mass: $X > Y = Z$. This question allowed us to verify that students were not simply giving a memorized response that drag forces are always equal to one another.

C. Data analysis

Student answers and reasoning for the scaffolding questions were coded as correct or incorrect. Two of the four scaffolding questions from stage 1 of the intervention were also used as screening questions [7,21]. If students indicated correctly, for the pancakes scenario, that object X experiences zero net force, and if they selected the correct free-body diagram for object X, they were deemed to possess the appropriate mindware, in other words, to have demonstrated knowledge sufficient to answer the pancakes target question correctly. Most of our analyses focused on the effects of our intervention on students who had already demonstrated some measure of conceptual understanding of Newton's second law by answering those two screening questions correctly. If there was a discrepancy between the multiple-choice option selected and the explanation provided, students who correctly stated that the forces must be balanced were considered to have demonstrated the appropriate understanding and were included in our study population. In other words, we allowed the explanation to override the actual multiple-choice response selected. The number of students giving such discrepant responses was typically small, no more than 5%. While we did not explicitly screen for the understanding that an object's weight near the surface of the Earth depends only on its mass (i.e., that two objects of the same mass also have the same weight), our experience as educators suggests that this is not a problematic concept for most students.

All target questions (pre, post, and transfer) were coded as to whether students had provided a correct answer with correct reasoning. Reasoning was coded as correct if a student explicitly gave a Newton's second law argument ("They both have the same force of gravity, so when they reach their terminal speeds, the air resistance is equal magnitude to their gravity forces and therefore equal to each other.") or stated that the masses were the same (thereby implying that the drag forces must therefore be the same—"They're the same mass"). On later questions, if students explicitly recognized that a scenario was the same as an earlier scenario ("it's the same as the last one"), they were coded as having used correct reasoning if they had done so previously. For many analyses, student responses to the target questions were treated as binary response variables (coded as "1" if a student responded correctly

with correct reasoning, and "0" if a student responded incorrectly or gave explicitly incorrect reasoning or no reasoning at all in support of a correct answer). The normal approximation to the binary distribution was used to calculate 95% Wald confidence intervals for the target questions [37]. McNemar's test [38] with a continuity correction [39] was used to check the statistical significance of shifts in performance from the pancakes target question (pretest) to the blobs target question (post-test) and the three balls target question (transfer task). McNemar's test is a nonparametric test for changes in the proportion of paired dichotomous nominal data [40]. Fisher's exact test was used to determine whether the correct response rate on any single task (e.g., the pancakes target question) differed between populations (e.g., between the students in the control condition and those in the treatment condition). Fisher's exact test is used as an alternative to the chi-square test when N values are lower because it provides an exact measure of the probability rather than an approximation [37].

V. RESULTS

A. Performance in stage 1

The majority of students, more than 88%, responded correctly to the two screening questions in the pancakes scenario. These students recognized that the net force on object X must be zero and its free-body diagram must show two balanced forces when object X is falling at a constant terminal speed. As shown in Table I, however, most students did not apply this understanding to arrive at a correct answer for the pancakes target question. The correct response rate for the pancakes question is less than 50%, much lower than for the screening questions. At WWU, response distributions for the pancakes target question were nearly identical regardless of whether or not students had previously demonstrated a correct understanding of the two screening questions. As described in Sec. III, many students' explanations for the common incorrect response were very brief; for example, one student wrote, "The drag force is greater on object Y because there is a greater surface area." A few students provided somewhat more detail, supporting their response with the drag force equation: "The equation for drag force $F = 0.5C\rho Av^2$ accounts for the cross-sectional area of the object, and object Y has a greater radius than object X so it has a bigger area and a greater magnitude of air resistance." Even though this student had previously indicated that object Y has a smaller terminal speed than object X (as did about 75% of students overall), the student did not reference that in their air drag explanation.

B. Comparisons in performance between treatment and control groups

The percentage of students responding correctly to each of the target questions (pre, post, and transfer), among those who had answered the screening questions correctly, is shown in Table II for both the treatment and control groups. These data are also presented visually in Fig. 10. A two-sided Fisher's exact test was used to test for differences in correct response rates (with correct reasoning) between the control and treatment groups on each of these questions. The effect size was measured using Cramer's V. Performance on the pancakes target question (before students had received any differentiated instruction) did not differ significantly between the control group and the treatment group ($p = 0.295$, $V = 0.078$), consistent with the random assignment of students from the same course to each group. Differences were also not significant between the two groups on the analogous blobs post-test target question ($p = 0.065$, $V = 0.140$) after differentiated instruction. However, on the near-transfer post-test question (the three-ball target question), the performance of the

TABLE I. Response distribution on the pancakes target question (pretest), among all responses and broken out by whether or not students had responded correctly on the screening questions. Values may not sum appropriately due to rounding. Data from UM will be discussed in Sec. VI.

	% of students giving response					
	WWU			UM		
	All students ($N = 228$)	Students correct on screening questions ($N = 201$)	Students incorrect on screening questions ($N = 27$)	All students ($N = 193$)	Students correct on screening questions ($N = 147$)	Students incorrect on screening questions ($N = 46$)
$F_{\text{drag on Y}} = F_{\text{drag on X}}$ (correct)	36%	35%	37%	28%	34%	9%
With correct reasoning	32%	33%	30%	23%	29%	5%
With incorrect or incomplete reasoning	3%	2%	7%	5%	5%	5%
$F_{\text{drag on Y}} > F_{\text{drag on X}}$	61%	61%	63%	67%	62%	84%
$F_{\text{drag on Y}} < F_{\text{drag on X}}$	4%	4%	0%	4%	7%	3%

TABLE II. Percentage of students responding correctly with correct reasoning on each of the target questions, among students who had responded correctly on the two screening questions. Values in parentheses represent 95% Wald confidence intervals. *P* values were calculated using Fisher's exact test. Cramer's *V* provides a measure of effect size. Shifts in correct response rate for each group are also shown, as are the differences in the shift between the control and the treatment group. Data from UM will be discussed in Sec. VI.

	WWU				UM			
	Control group ($N = 98$)	Treatment group ($N = 102$)	p	V	Control group ($N = 72$)	Treatment group ($N = 74$)	p	V
Pancakes target (pretest)	37% (27%–46%)	29% (21%–38%)	0.295	0.078	35% (24%–46%)	27% (17%–37%)	0.371	0.083
Blobs target (post-test)	77% (68%–85%)	87% (81%–94%)	0.065	0.14	40% (29%–52%)	57% (45%–68%)	0.049	0.165
Three balls target (transfer)	61% (52%–71%)	76% (68%–85%)	0.022	0.165	39% (28%–50%)	51% (40%–63%)	0.139	0.125
Shift: pre to post	40% (29%–50%)	58% (48%–68%)			6% (−5%–16%)	30% (17%–43%)		
Shift: pre to transfer	25% (12%–37%)	47% (35%–59%)			4% (−7%–15%)	24% (10%–38%)		
Difference in shifts: pre to post	18% (4%–33%)				24% (8%–40%)			
Difference in shifts: pre to transfer	23% (6%–40%)				20% (2%–38%)			

treatment group on the question was significantly better than the performance of the control group ($p = 0.022$, $V = 0.165$), with a small effect size.

C. Shifts in performance from pretest to assessment tasks

Student performance did appear to improve overall as a result of the instruction present in both conditions of the intervention. Four McNemar's tests were conducted to assess the performance on the blobs and three balls target questions in comparison to the performance on the pancakes target question for each group. The results, shown in Table III, indicate that for each group, the observed shift in correct response rates from the pretest to the post-test and from the pretest to the transfer task was statistically significant.

As an additional measure of the effect size associated with the shifts in performance, confidence intervals for the shifts in performance from pretest to post-test and from pretest to transfer were constructed. These confidence intervals, calculated by treating the responses on the pancakes target question, the blobs target question, and the three-ball target question as dependent samples, are shown in rows 4 and 5 of Table II. The 95% C.I. for the shift from pre- to post- and from pre- to transfer does not include zero for either the control or treatment group, confirming that the improvements in performance after stage 2 instruction were statistically significant for both groups.

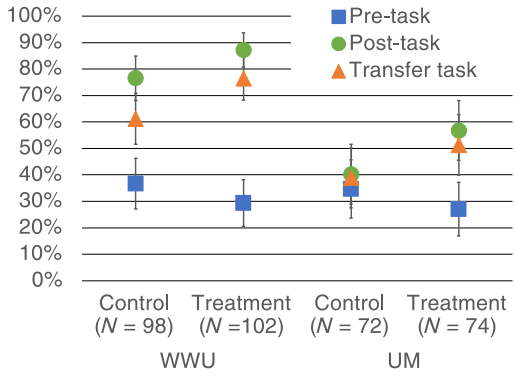


FIG. 10. Performance on the target tasks before (pre) and after (post and transfer) differentiated instruction. Error bars represent 95% Wald confidence intervals. Data from UM will be discussed in Sec. VI.

The improvements in performance were greater for the treatment group than for the control group for both the blobs question and the three-ball question. As shown in the bottom two rows of Table II, the difference between the two groups in the shift from pretest to post-test, calculated by treating each group as an independent sample, was 18% (95% Wald C.I.: 4%–33%). The difference in shifts from pretest to transfer question was 23% (95% Wald C.I.: 6%–40%). In other words, the improvement in performance from pretest to post-test is highly likely to be between 4% and 33% greater for the treatment group than for the control group. From pretest to transfer question, the improvement in performance was between 6% and 40% greater for the treatment group than for the control group.

VI. REPLICATION EFFORTS

A. Context for replication

In order to gauge the replicability of the results, the experiment was repeated in the calculus-based introductory mechanics class at the University of Maine (UM). This course was similar to Western Washington in many respects, including the student population and the use of *Tutorials in Introductory Physics*. Students in the course at UM, however, had received only minimal instruction on air resistance. The lecture portion of the course devoted a

TABLE III. Test statistics (z_0^2), p values, and Cohen's g (a measure of effect size; values about 0.25 are considered a large effect) for each of the McNemar's tests that were conducted on the target question data. Data from UM will be discussed in Sec. VI.

		WWU		UM	
		Control group ($N = 98$)	Treatment group ($N = 102$)	Control group ($N = 72$)	Treatment group ($N = 74$)
Test: Post vs pre	z_0^2	35.4	57.0	1.14	16.1
	p	< 0.001	< 0.001	0.143	< 0.001
	Cohen's g	0.45	0.48	0.14	0.36
Test: Transfer vs pre	z_0^2	13.1	42.3	0.53	10.1
	p	< 0.001	< 0.001	0.233	< 0.001
	Cohen's g	0.27	0.4	0.08	0.28

portion of a single lecture to the topic of air resistance, and students had not completed any labs on air resistance. We thus expected the mindware regarding air resistance to be less well developed in the UM students. In the UM context, students completed the instructional and assessment sequence as part of a weekly participation-based homework assignment. The assignment was given after all instruction on forces, including air resistance, had concluded. Students received participation credit for completing the assignment, regardless of the correctness of their responses. As at WWU, the homework assignment was delivered via Qualtrics, and students from the single lecture section were randomly assigned by the Qualtrics software into treatment and control conditions.

B. Replication results

Results from the administration of the experimental sequence are shown in Tables I–III, and Fig. 10. Overall, correct response rates across all tasks at UM were somewhat lower than at WWU, consistent with students at UM receiving less instruction on the target topic than students at WWU. Just under 75% of students at UM responded correctly to both screening questions (compared with 88% at WWU). Among those students who did not respond correctly to the screening questions, very few were successful at answering the pancakes target question (pretest). Among those who did respond correctly to the two screening questions (the primary population of interest in our study), however, the response distribution on the pancakes target question (pretest) was remarkably similar to the distribution at WWU.

The difference between the UM and WWU populations is especially evident in the post-test and transfer tasks, on which UM students performed substantially less well than WWU students. In fact, for the UM students in the control group, there was no significant shift in performance from pretest to either post-test or transfer task. The 95% CI for each shift includes zero (rows 4 and 5 of Table II), and the McNemar test statistics for the control group at UM are nonsignificant (Table III). Consistent with the results from WWU, however, the treatment group at UM demonstrated improvements in performance on the target task from both pretest to post-test and pretest to transfer task. Both shifts in performance are significantly different from zero, as shown in Table II, and McNemar's tests yielded significant results with large effect size, as shown in Table III.

VII. DISCUSSION

We have found that the pancakes target question is difficult for students. On the screening questions, many students demonstrated the availability of knowledge and skills that would seem sufficient to answer the target question correctly and yet they did not apply it successfully to the target question. Comparing results from WWU and UM demonstrates that increased instruction on air resistance did not result in substantially higher correct response rates on the pancakes target question in stage 1. Similar phenomena have been observed in other physics contexts [6–9,21]. As in those studies, here we attribute the inconsistencies in student responses to the nature of human reasoning itself [10,12,14]. The high prevalence of the incorrect response that object Y experiences a greater drag force than object X may be due primarily to the high salience of the cross-sectional areas of the two pancake-shaped objects. This salience may lead many students to generate an intuitive model along the lines of "objects of larger cross-sectional area experience greater air resistance." This model, while sensible and related to potentially productive lines of reasoning, will not lead to the normative answer for the pancakes question involving objects moving at terminal speeds. An active mental model "objects of larger cross-sectional area experience greater air resistance" that leads directly to the output of the answer that $F_{\text{drag on Y}} > F_{\text{drag on X}}$ without further review is consistent with the reasoning pathway involving hazards A and B in Fig. 1. The brevity of the explanations for this common incorrect response suggests that indeed, for many students, the analytic process did not engage to evaluate the model leading to the response.

Even for those students whose analytic process did engage, the analysis may have been influenced by confirmation bias or other cognitive biases. As noted, when students did provide lengthier explanations, many drew upon the drag force equation $F_{\text{drag}} = \frac{1}{2}C\rho A v^2$. A cursory inspection of this equation may provide a quick confirmation of the student's initial model, without recognition that the difference in terminal speeds of the two objects means that the equation, on its own, is inconclusive for comparing the drag forces. In this reasoning pathway, the validity of the initial answer has not been scrutinized through an active search for alternative models. Process 2 has engaged in a shallow, rather than a sustained manner, consistent with the reasoning pathway involving hazard C. Overall, we find that students' performance on the pancakes question is thus fully consistent with a dual-process model for human reasoning.

In the experiment, we used the context of objects falling at terminal speed to test the effects of two different instructional approaches. One was designed simply to provide repeated, scaffolded practice with the normative mindware (the control condition)—in effect supporting process 1 and targeting hazard A. The other not only allowed students to practice the normative mindware but also guided them to reflect upon and analyze their intuitive response (the treatment condition)—helping to activate process 2 and supporting its sustained engagement, thereby targeting hazards B and C. Among students who had already demonstrated a basic understanding of the necessary concepts, the treatment group outperformed the control group across several measures at both institutions. At both WWU and UM, the treatment group demonstrated larger gains in performance from the pretest task to the post-test and transfer tasks, and at WWU, the treatment group also demonstrated better absolute performance on the transfer task. This suggests that for some students, instruction that targets hazard C by supporting the analytic process in evaluating a first-available mental model will be more useful than instruction that targets hazard A by reinforcing the normative mindware.

While the approach supporting process 2 was more effective than the approach supporting process 1, some students in the control group did derive benefits from the repeated practice with mindware in stage 2. Before attempting the assignment used for our study, WWU students had received much more research-based instruction on air resistance (including a lab focused on concept development) than UM students. As shown in Table II and Fig. 10, the performance of the WWU students in both the treatment and control groups on the assessment questions (both post-test and transfer) was substantially better than on the pancakes pretest question, whereas at UM only students in the treatment group demonstrated improved performance. As noted in Fig. 4 and described in Sec. IV B 2, in stage 2 of the assignment, all students experienced instruction that included an explicit discussion of the correct answer and reasoning for the pancakes target question. The blobs target question, which served as a post-test, is completely analogous to the pancakes target question, differing only in surface features. It appears that students at WWU had been primed sufficiently for even students in the control group to be able to apply this reasoning in a new setting. This may be because students at WWU had developed their understanding of air resistance to the point that the brief explanation provided to the control group was enough to "raise a red flag" and then allow them to successfully analyze and reject the incorrect intuitive model. At UM, however, where students had experienced less instruction on air resistance, even if the students were alerted to an error in their response, students may not have had sufficient available and relevant mindware to successfully analyze the intuitive model on their own. When provided with the supports for analysis present in the treatment condition, however, they were more likely to be successful.

Although in this study the instruction that emphasizes the normative reasoning was not as successful as the instruction that supported students in the analysis of the common incorrect model, it is likely that with much more practice of the normative mindware, spanning a longer time period, the performance of all students would improve even without guided reconciliation of the first-available model with the correct reasoning. Eventually, the normative reasoning will have been practiced to the point of automaticity [15,41], such that it becomes the default model produced by the intuitive process. In other words, with sufficient practice, the accessibility [11,22] of the normative reasoning will increase to the point where a reasoner's intuitive mental model is itself normative and leads to the correct response. This is the case for many physics experts: When confronted with an unfamiliar situation like the pancakes target question, the default model may be anchored to fundamental principles like Newton's second law rather than to context-dependent formulas like $F_{\text{drag}} = \frac{1}{2}C\rho A v^2$. Even if experts are distracted by the high salience of the difference in cross-sectional areas and initially consider the common incorrect model, they are also more likely to check their reasoning by testing any provisional models against Newton's second law. The process of developing an expertlike intuition, however, is time-consuming and resource-intensive—prohibitively so for many introductory physics courses. Thus, instruction that supports students in engaging the analytic process to evaluate their likely first-available model, and provides opportunities for students to reconcile their first-available model with the normative physics reasoning, can be more efficient at helping students arrive at a correct answer. A long-term goal of such instruction would be to help students learn to engage their analytic process more productively in completely novel situations, an expertlike behavior and an outcome that is not likely to happen in the presence of instruction that only trains mindware.

VIII. CONCLUSIONS AND FUTURE WORK

In this work, we have provided one example demonstrating that for students who have already demonstrated the availability of the necessary concepts, instruction that supports these students in analyzing their likely intuitive mental model leads to greater learning benefits in the short term than instruction that focuses solely on providing practice with the normative mindware. We interpret these results through the lens of dual-process theories of reasoning to suggest that once students have a grasp of the relevant physics, focusing instruction on supporting the productive engagement of the analytical process may be more efficient and effective than attempting to provide enough practice to automatize the correct reasoning.

The work presented here took place in one single short online assignment, on one single topic that is frequently skipped in introductory physics courses. The median completion time for the entire instructional and assessment sequence was around 24 min for both control and treatment groups, around half the duration of a typical tutorial or recitation session. Whether the difference in performance between the treatment and control groups persisted beyond the length of the assignment remains an open question that is currently under study. Similarly, whether this intervention strategy could be adapted to other topics in physics is the subject of ongoing investigation. Other instructional strategies that leverage dual-process theories in different ways for other student populations, including those with less well-developed initial mindware, are also under ongoing development and assessment.

The development of a solid foundation of conceptual understanding in physics remains an important goal of physics instruction. However, the results presented here suggest that attending to the development of students' reasoning skills, and in particular, helping them develop the capacity for analyzing their own thinking, is a critical component for improving performance. Designers of research-based curricula have long incorporated instruction designed to address specific difficulties that students encounter when learning physics content [2]. In this work, we demonstrate an example in which a persistent difficulty may in fact be explained using a dual-process framework for human cognition. Consistent with that framework, we have described a specific strategy for addressing difficulties by providing students with the opportunity to reconcile their initial (incorrect) mental model with the normative model and then prompting reflection on the relative usefulness of the normative model and the common incorrect model. While this instructional approach shares many similarities with *elicit-confront-resolve* or *refining intuitions,* drawing on dual-process theories allows us to gain insight into the mechanisms by which the instruction—and by extension other successful research-based instructional materials—may be successful.

We hope that the educational experience described in this work is a positive one for students, despite the inherent invitation to grapple with a common error. We view the treatment condition, in particular, as providing an example of "helping physics students learn how to learn" [42]. In the reconciling process that is central to the treatment, students redirect the common intuitive idea, that larger size corresponds to greater drag force, toward a more productive line of reasoning, wherein students recognize and explain that not just the size but also the speed differs from one object to the next. Rather than demonstrating to students that their intuition does not apply to physics, the treatment thus guides students to productively apply their intuition in a way that is consistent with other valid knowledge. We ultimately hope that, as part of a coherent approach that foregrounds physics reasoning and decision making through the lens of dual-process theories, we may begin to normalize the making of mistakes and promote a culture of thoughtful reflection. However, our study does not measure students' epistemologies or other aspects of their affective responses to instruction. Measurements of such affective characteristics could be a valuable avenue for exploration in future research.

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, and No. DUE-1821561. Thanks are due to Paula Heron for her contributions during many discussions of this work.

[1] P. R. L. Heron, Empirical investigations of learning, and teaching, part I: Examining, and interpreting student thinking, in *Proceedings of the International School of Physics "Enrico Fermi" Course CLVI: Research on Physics Education*, edited by E. F. Redish and M. Vincentini (IOS Press, Varenna, Italy, 2003), pp. 341–350.

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

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

[4] D. Hammer and A. Elby, Tapping epistemological resources for learning physics, J. Learn. Sci. **12**, 53 (2003).

[5] B. A. Lindsey, M. R. Stetzer, J. C. Speirs, W. N. Ferm, and A. van Hulten, Investigating student ability to follow reasoning chains: The role of conceptual understanding, Phys. Rev. Phys. Educ. Res. **19**, 010128 (2023).

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

[7] 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).

[8] M. Kryjevskaia and N. Grosz, Examining students reasoning in physics through the lens of the dual process theories of reasoning: The context of forces and Newton's laws, in *Research and Innovation in Physics Education: Two Sides of the Same Coin*, edited by J. Guisasola and K. Zuza (Springer, Cham, Switzerland, 2020), pp. 91–108, 10.1007/978-3-030-51182-1_8.

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

[10] A. F. Heckler, The ubiquitous patterns of incorrect answers to science questions: The Role of Automatic, Bottom-up Processes, in *Psychology of Learning and Motivation—Advances in Research and Theory* (Elsevier Inc., Cambridge, MA, 2011), pp. 227–267, 10.1016/B978-0-12-387691-1.00008-9.

[11] 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).

[12] M. Kryjevskaia, P. R. L. Heron, and A. F. Heckler, Intuitive or rational? Students and experts need to be both, Phys. Today **74** (8), 28 (2021).

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

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

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

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

[17] M. Mays, M. R. Stetzer, and B. A. Lindsey, Supporting student construction of alternative lines of reasoning, presented at PER Conf. 2021, virtual conference, 10.1119/perc.2021.pr.Mays.

[18] T. Fittswood, D. J. Rosen, and M. R. Stetzer, Insights from an intervention designed to support consistent reasoning, presented at PER Conf. 2022, Grand Rapids, MI, 10.1119/perc.2022.pr.Fittswood.

[19] 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).

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

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

[22] D. Kahneman, A perspective on judgment and choice: Mapping bounded rationality, Am. Psychol. **58**, 697 (2003).

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

[24] M. Kryjevskaia, M. R. Stetzer, and T. K. Le, Failure to engage: Examining the impact of metacognitive interventions on persistent intuitive reasoning approaches, presented at PER Conf. 2014, Minneapolis, MN, 10.1119/perc.2014.pr.032.

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

[26] V. A. Thompson, Dual-process theories: A metacognitive perspective, in *In Two Minds: Dual Processes, and Beyond*, edited by K. Frankish and J. S. B. T. Evans (Oxford University Press, Inc., Oxford, UK, 2012), pp. 171–195, 10.1093/acprof:oso/9780199230167.003.0008.

[27] D. N. Perkins, *Outsmarting IQ: The Emerging Science of Learnable Intelligence* (Free Press, New York, 1995).

[28] H. A. Simon, What is an explanation of behavior?, Psychol. Sci. **3**, 150 (1992).

[29] M. Nieberding and A. F. Heckler, Evolution of response time and accuracy on online mastery practice assignments for introductory physics students, Phys. Rev. Phys. Educ. Res. **19**, 020111 (2023).

[30] E. Etkina and G. Planinsic, Thinking like a scientist, Phys. World **27**, 48 (2014).

[31] A. Boudreaux, E. Dibeh, E. Moran, and B. Lindsey, Empirical investigation of student understanding of terminal speed motion of falling objects (to be published).

[32] A. Boudreaux and B. Lindsey, Investigation of student reasoning about air resistance and terminal speed behavior of falling objects, in *Proceedings of the APS April Meeting* (2023).

[33] 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).

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

[35] K. Heller and P. Heller, *Cooperative Problem Solving in Physics. A User' s Manual* (University of Minnesota, Minneapolis, MN, 2010), p. 310.

[36] Qualtrics, http://www.qualtrics.com (2021).
[37] A. Agresti, *Categorical Data Analysis*, 3rd ed. (John Wiley & Sons, Hoboken, NJ, 2013).
[38] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, Psychometrika **12**, 153 (1947).
[39] A. L. Edwards, Note on the "correction for continuity" in testing the significance of the difference between correlated proportions, Psychometrika **13**, 185 (1948).
[40] M. A. Morrison, McNemar's test, in *Encyclopedia of Research Design* (SAGE Publications, Inc., Thousand Oaks, CA, 2012), p. 780, 10.4135/9781412961288.n235.
[41] 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).
[42] A. Elby, Helping physics students learn how to learn, Am J Phys **69**, S54 (2001).