Helping Students Recognize and Resolve Reasoning Inconsistencies: An Application of Dual-Process Theories

Kristin Kellar

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Paula Heron, Chair

Peter Shaffer

Suzanne White Brahmia

Program Authorized to Offer Degree:

Physics

University of Washington

Abstract

Helping Students Recognize and Resolve Reasoning Inconsistencies: An Application of Dual-Process Theories

Kristin Kellar

Chair of the Supervisory Committee:

Paula Heron

Department of Physics

Recent educational objectives have focused on preparing students for the future workforce and society through the development of 21st-century skills. Physics education research (PER), in combination with this broad educational objective, has informed discipline-specific educational goals, where one emphasis lies in helping students correctly apply physical concepts in a variety of contexts, including those that are unfamiliar. It has been observed that some students demonstrate reasoning inconsistencies on such tasks that tend to elicit an incorrect intuitive idea likely due to contextual surface features and low-level cognitive processes. This dissertation explores how the theoretical framework of dual-process theories (DPTs) can give insight into this type of student inconsistency and inform the development of intervention strategies ultimately intended to help students succeed on these kinds of tasks in physics.

Dual-process theories of reasoning posit that humans reason using two processes: process 1 is intuitive and automatic while process 2 is analytical and deliberate. When confronted with a situation that requires a decision, process 1 automatically engages to produce a

provisional model to address the task at hand. The reasoner may or may not activate process 2 to assess the validity of this provisional model before arriving at a conclusion. In the case where a student with sufficient content knowledge fails to correctly answer a physics problem, process 1 may have generated an incorrect intuitive model which process 2 did not override. In this dissertation, research is presented on the use of DPTs to develop and evaluate sets of questions intended to (1) disentangle students' conceptual understanding from their reasoning approaches and (2) intervene to support students' process 2 engagement on tasks that often prompt an incorrect process 1 response. Two question sequences are described and investigated, one in the context of a pulse on a spring and one in the context of boxes at rest on rough surfaces. The results suggest that the sequences fulfill the two intentions described above: they serve to distinguish errors that stem from overreliance on process 1, and they help students who make those errors to engage process 2 and ultimately reach correct conclusions. This work serves as a model for developing question sequences in other contexts that, in combination, may have more general long-term and far-transfer implications on student performance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I count myself blessed for all the support that has surrounded me as I have worked toward the completion of this dissertation. I want to thank the many mentors, colleagues, family, and friends who have provided guidance, encouragement, and support to me on this journey. First, I would like to acknowledge Paula Heron for her tremendous support as my mentor. Paula, thank you for your guidance and example of service. You have been a significant influence in the development of my research, writing, and presentation skills. I am so grateful for the way you have prioritized my success. You were always available and willing to provide feedback, support, and flexibility.

I would also like to thank Peter Shaffer and Suzanne White Brahmia for their time and dedication in providing feedback during the writing and presentation of this work. Peter, I also appreciate your help obtaining various data necessary in the analysis involved in this research. Additional thanks are due to the remaining members of my committee: Barbara Wakimoto, Silas Beane, and Kai-Mei Fu. Thank you for your time and consideration in evaluating my work. In addition, I recognize all the contributions of those associated with the University of Washington (UW) Physics Education Group (PEG), past and present. Donna Messina, Anne Alesandrini, John Goldak, Cam Flynn, Al Snow, Jack Chapman, Ella Henry, Charlotte Zimmerman, Jared Canright, Lisa Goodhew, Dean Bretland, Bert Xue, Sheh Lit Chang, Alexis Olsho, Jesse Ashworth, Taylor GurrEithun, Qirui Guo, Yasmene Elhady, Ellie Chew, Ava Aflatooni, Alex Reynolds, Sabrina Cheng, Lauren Bauman, Rachel Scherr, and Clausell Mathis have been a wonderful community, always willing to give feedback and support.

Additionally, I would like to acknowledge the cooperation of all the instructors who allowed me to modify course materials and collect data from their courses. In particular, Peter Shaffer, Nikolai Tolich, Kazumi Tolich, and David Smith were consistently willing to grant me access to relevant instructional platforms for the courses and permitted me to modify online quizzes to collect student response data. I also appreciate the aid of Peter Shaffer in teaching me how to use the instructional platforms to modify and administer the quizzes.

Not only have I been surrounded by a collaborative group at the UW, but within the physics education research (PER) community. I would like to thank Mila Kryjevskaia, Mac Stetzer, Andrew Boudreaux, Beth Lindsey, and Drew Rosen from our dual-process theories of reasoning collaboration for their support in conducting research. They were always available to discuss experimental design, intervention development, interpretation of results, recommendations of resources, and communication of results. I am also thankful for the many conversations I have had with others in the PER community including Andrew Heckler—who provided valuable data analysis recommendations—and Caleb Speirs, Mikayla Mays, Safana Ismael, and Thomas Fittswood, with whom I have had stimulating conversations about dual-process theories in PER.

I wouldn't have pursued physics education research without many others in my life who have been by my side. Mary Kemp, thank you for sparking my interest in physics in high school. You made physics fun and understandable and inspired me to go into physics education. Jess Dowdy, thank you for providing opportunities for me to get involved in STEM education and for advocating for me to pursue the teaching of physics. I would also like to acknowledge Andrew Huddleston. I appreciate your time, support, and mentorship as you encouraged me to pursue my newfound passion of physics education research while learning and practicing education research at Abilene Christian University (ACU). I also appreciate Gary Strickland for allowing me the opportunity to learn the art and practice of teaching while collaborating with me on my first teacher action research project. A special thank you goes out to my

former colleagues at Talkington School for Young Women Leaders. It was a privilege to teach alongside such inspiring educators. I would especially like to thank Catherine VanHoorebeke for her guidance and encouragement as I developed my teaching skills.

I would not be where I am today without the emotional, mental, and tangible support of family and friends. Mom and Dad, I appreciate your belief in me, all your encouragement and support on this journey, and letting me go off to pursue my passion. Justin Kellar, thank you for your willingness to support the pursuit of my dream in physics education research, for the time you sacrificed to pick up the slack while I worked countless hours, and for your continued love and encouragement. Claire, thank you for your patience while I completed this work, for sharing in the excitement of my accomplishments, and your unconditional love. Holland and Laura, words cannot express how much I appreciate your continual support in providing childcare, which allowed me the time and focus necessary to complete this research. I am also thankful for your service in caring for my needs and your willingness to give of yourselves. Last but not least, I am grateful for my church family, who has been so understanding while I committed my time to this endeavor. Thank you for all your words of encouragement and prayers along the way.

This material was 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. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

DEDICATION

To my family, who was a constant source of support and encouragement

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Current educational goals focus on helping students develop "21st-century skills," which involve cognitive, intrapersonal, and interpersonal competence valuable in preparing for the future workforce and society [1]. This broad educational goal of developing 21st-century skills along with the physics education research (PER) literature has informed the development of a set of discipline-specific goals proposed for physics education [2].

One such goal related to cognitive competency is to advance student expertise in physics by helping them "develop well integrated knowledge structures in order to achieve deep learning in physics" [2]. Without well-organized knowledge structures, novices often rely on surface features of a given problem [3, 4, 5] and are limited in their ability to apply conceptual approaches to novel situations [2]. Experts, on the other hand, are characterized as having knowledge organized around core physics principles that are useful in addressing new and unfamiliar situations [6, 7, 8, 9]. Students' cognitive development occurs in the transition from fragmented and context-dependent knowledge structures to integrated and less context-specific structures [2].

Another major goal within the cognitive domain is to foster scientific reasoning [2], which often refers to a set of skills that support critical thinking, problem solving, and creativity in STEM education [10]. Students' development of scientific reasoning can lead to the improvement of 21st-century skills such as critical thinking and decision making [10] and has been observed to promote success on novel reasoning tasks [11].

A common theme of the goals for physics education outlined above is helping students transition from novices to experts, with a specific intent to progress their ability to apply physical concepts correctly on a variety of tasks including those involving new and unfamiliar contexts. In order to achieve this, more research into effective educational approaches is needed [2]. Reform in science education towards this goal has been founded on a scientific basis of learning, developed from an understanding of human cognition in such areas as the structure of knowledge as well as reasoning [9].

Below I highlight a classroom example found in the literature that demonstrates inconsistent response patterns associated with novice behavior in students. Then I provide a brief list of possible perspectives for interpreting such results, including those related to knowledge structures and low-level cognitive processes. In the section that follows, I describe a theoretical framework from cognitive science (used in this dissertation) that can be leveraged in instruction to help promote more expert-like behavior in students.

In the kinematics graph task shown in Fig. 1.1, a position-time graph shows the motion of two different cars, and students are asked to determine when the speeds of the two cars are the same. This requires application of the concept that the slope of a position-time graph determines velocity, and so the point on the graph where the two lines have the same slope, time A, is where the two cars have the same speed. When given to introductory undergraduate physics students following relevant instruction, only 60% of students correctly chose time A while 40% of students chose time B where the lines cross [12]. One may postulate that the students who chose time B have a misconception (incorrect idea or framework that appears to be established, coherent, and context-independent [13]) that the height on a position-time graph determines speed.

This may be true for some students; however, responses on a related task given to the same students were surprisingly inconsistent. Fig. 1.2 shows the related task involving another position-time graph, about which they were again asked to determine when the speeds of the two cars were the same. Almost all students correctly identified that both cars have the same

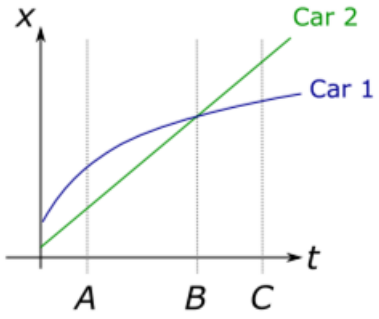


Figure 1.1: Diagram associated with a kinematics graph task. Students are asked to determine when the speeds of the two cars are the same. About 60% of students correctly chose time A while the other 40% of students chose time B [12, 13, 14].

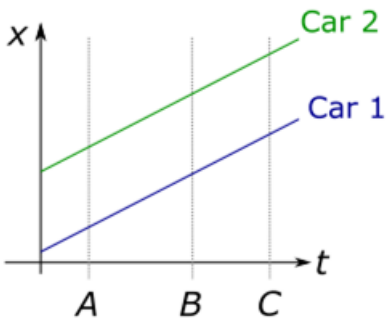


Figure 1.2: Diagram associated with a kinematics graph task. Students are asked to determine when the speeds of the two cars are the same. Almost all students correctly responded that both cars have the same speed at all times [12, 13, 14].

speed at all times [12]. If some students held a stable, context-independent misconception that height determined velocity, a greater percentage of students would have responded that, based on this graph, the cars were never traveling at the same speed. It seems that some students were able to apply the correct content knowledge in one context, but not in another.

One lens for interpreting this response pattern is through the framework of "resources." Resources are defined as small cognitive structures that are activated momentarily in context-sensitive ways [15, 16, 17]. Examples include the ideas "closer means stronger" and "main-

taining agency" (effort must be continued to maintain an effect) [15]. A particular resource may be accessed in one context but not another, and different resources may be activated in sets [15]. From this perspective, students may have relied on a different set of resources in response to each graphing task based on the context. However, without further study, it is unclear what resources are commonly activated for these tasks and what mechanisms predict the activation of particular resources. Let us postulate some low-level cognitive mechanisms that may be at play.

One possible mechanism related to the context involves the features of each graph that seem to capture the most attention. In the case of the first graph in Fig. 1.1, the point of intersection, while irrelevant to the task at hand, may be a particularly salient feature of the graph for some students. These students may have relied on this feature in their response. In contrast, the graph in Fig. 1.2 contains lines that have the same slope everywhere, which might have called attention to the relevant feature of slope and may have contributed to higher performance on this question.

Another low-level mechanism that may be contributing to incorrect responses to the task shown in Fig. 1.1 is related to relative processing time between the dimensions of height and slope on a graph. Students have been observed to compare the heights of two points on a graph significantly faster than the slopes of two points [18]. Additionally, in contexts where slope is the correct dimension to compare, some students appear to know the correct answer yet respond incorrectly because of a preference to answer quickly even when no time limit is imposed [18]. On the task shown in Fig. 1.1, students may be more inclined to choose point B, which focuses on the dimension of height rather than point A, which compares the slopes of the two functions, due to the nature of relative processing time for height and slope.

Attentional allocation and relative processing time are examples of mechanisms of "bottom-up processes" that may contribute to inconsistent student response patterns on various tasks. Bottom-up cognitive processes are characterized as basic, unconscious, and automatic [14, 16, 17, 18]. In contrast, top-down processes consist of higher-level structures

or processes, taking the form of conceptual understanding or explicit reasoning [14]. The results from student responses to the graph tasks discussed above suggest that student reasoning cannot be solely understood as an application of top-down mental processes like the application of previously learned conceptions, but bottom-up processes as well which can be characterized by lower-level mental processes including perceptions built from sensory input like context cues [14]. Therefore, a theoretical framework encompassing both types of cognitive processes can be useful in gaining insight into student inconsistencies and provide a foundation for instructional approaches intended to help students recognize and overcome such inconsistencies as part of their development towards expert-like behavior in physics.

1.2 Theory

Student response patterns, similar to those observed on the graph tasks above, can be interpreted from the perspective of dual-process theories of reasoning (DPTs) [19]. There are many theories that fall into this category, but they all generally agree that humans reason using two processes. Process 1 is automatic and intuitive (activated in all situations) while process 2 is more deliberate and rule-based (only engaged in some situations).

The basic tendency of humans is to rely on processing mechanisms that require low computational expense associated with process 1 [20]. However, process 1 is not always reliable, and process 2 is necessary to arrive at judgments that are appropriately reasoned. For example, consider the following question: "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?" [21]. Chances are that the first thought that came to mind (generated by process 1) as a solution to this problem was ten cents. Upon further reflection (through the engagement of process 2 thinking), it can be determined that the accurate solution to this problem is five cents. Even when engaging process 2, the reasoning process is subject to errors like confirmation bias and weak knowledge [22]. Confirmation bias entails rationalizing a model instead of searching for alternative possibilities [23], and weak knowledge is characterized by insufficiency in one's conceptual framework for reasoning

in a particular context [22]. Therefore, even with the basic knowledge and skills needed to approach a reasoning task, reasoners can fall prey to various reasoning hazards that produce inaccurate conclusions.

1.3 Research Goals

Success in a physics course relies partly on the robustness of one's conceptual understanding as well as one's ability to effectively reason. The latter is especially relevant when it is desirable for physics students to be able to apply their knowledge to new and complex situations where their intuition may lead them astray. The aim of this research is to (1) validate screening-target question sequences (defined below) intended to disentangle students' reasoning approaches from their conceptual understanding, and (2) develop an intervention strategy intended to support students' process 2 thinking.

1.4 Goal (1) Validating Screening-Target Question Sequences

Given the focus of this research on human reasoning as an important factor in student success, it is helpful to not only gain insight into students' conceptual understanding of a particular topic but their potential reasoning pathways as well. Kryjevskaia *et al.* [24] developed question sequences intended to disentangle these two aspects of student approaches to answering physics problems. The question sequences consisted of "screening" and "target" questions within the context of specific physics concepts. Student performance on an initial screening question or questions provided information about their content knowledge. Then a target question was asked which required the same content knowledge but included a distracting feature that would likely trigger unproductive process 1 ideas. For students who demonstrated adequate content knowledge by answering the screening question correctly, a subsequent incorrect answer on the target question can be an indication of insufficient reasoning, not deficiency of conceptual understanding. These students presumably did not effectively engage process 2 thinking to override an initial incorrect idea with a correct conclusion on the target question.

Screening-target question sequences can prove a useful tool for understanding student inconsistencies from a human reasoning standpoint [25]. However, in developing these types of question sequences in various contexts, it is vital to establish if they support two main assumptions:

Assumption (1): The screening question accurately identifies students as having, or not having, adequate content knowledge necessary for successfully answering the target question.

Assumption (2): For students who correctly answer the screening question, success on the target question is largely a matter of cognitive reflection on an incorrect process 1 intuition.

Additionally, it is of interest whether dual-process theories are sufficient to explain the observed student response patterns. As such, the first goal of this research aims to provide a model for validating screening-target sequences by testing the two assumptions above while assessing the adequacy of DPTs interpretations.

1.5 Goal (2) Developing Intervention Strategies

Since human reasoning as described by DPTs has an impact on student performance on various tasks, developing intervention strategies that attend to student reasoning may be a useful tool for improving student success in physics. Research has been done to this very end by utilizing strategies such as additional instruction on easily accessible concepts, in-class hands-on activities, guided questions, individual and group work, and reasoning chain construction tasks in various physics contexts [24, 26, 27, 12, 28].

The research presented in this dissertation documents another effort to help students overcome reasoning inconsistencies, observed by their improved performance on tasks that tend to elicit an incorrect intuitive idea. It contributes to the library of DPTs interventions consisting of a variety of physics contexts, intervention tactics, and instructional strategies. The approach taken in the interventions developed as part of this dissertation consist of written sets of questions designed to steer students away from intuitively appealing unproductive ideas,

remind them of relevant knowledge, and support effective process 2 engagement through encouraging the consideration of alternative lines of reasoning. They are intended to provide a relatively short "micro-intervention" that does not significantly impose on established instruction yet seeks to aid students in correctly applying physics content. The purpose of this design is to allow for the development of such interventions in multiple contexts that can be embedded in a course without overburdening it. Research would need to be conducted to determine if regular incorporation of these DPTs-based interventions could have a long-term or far-transfer impact on student reasoning. However, regular incorporation of these interventions is outside the scope of this dissertation. The focus of this work is on the iterative development and analysis of the micro-intervention strategy as it is applied in two contexts.

One context involves a screening-target question sequence regarding a pulse on a spring adapted from [24] where the relevant content knowledge involves understanding what factors affect the speed and width of a pulse. The intervention was intended to support students' process 2 thinking on the target question and was evaluated to determine whether it seemed to cause students to recognize a need to reflect on an initial incorrect model and effectively engage in analytic thinking to arrive at a correct response. It also serves as a model that can be used to produce similar micro-interventions in other content areas as well.

The intervention strategy developed in the pulses context was then applied to a scenario involving boxes at rest on rough surfaces where a conceptual understanding of Newton's second law is relevant. It serves as a proof of concept for the DPTs-based intervention strategy and its evaluation in a new context.

1.6 Review of Prior Research

The research described in this dissertation, as it relates to goal (1), largely stems from work done by Kryjevskaia *et al.* in [24] who developed and implemented screening-target question sequences intended to disentangle students' conceptual understanding from reasoning approaches as discussed above. They applied this methodology to the context of a pulse on

a spring as well as in the context of capacitors connected in series.

The authors claim that observed student reasoning patterns were consistent with dual-process theories. Of students who answered the screening questions correctly on each question sequence, only a fraction of them went on to answer the target question correctly, and the majority of those who answered the target question incorrectly reasoned using intuitive ideas thought to be generated by process 1. While it appears that DPTs have explanatory power for the response patterns observed, this interpretation assumes that the screening questions were accurately identifying students with the appropriate content knowledge, and subsequently, that those with conceptual understanding who answered the target question incorrectly were in fact relying on a misleading process 1 model of thinking. Chapter 2 of this dissertation investigates these assumptions with regard to the pulse-on-a-spring question sequence reported in [24].

Goal (2) of this dissertation, to develop intervention strategies that address student reasoning, has been a common goal of researchers in the field of physics education research focused on dual-process theories of reasoning [26, 27, 12, 28]. Two main approaches are typically used to guide student success on tasks that often elicit unproductive intuitions:

(a) increasing the likelihood of process 1 generating a correct model (*e.g.* through repeated practice), and

(b) increasing the likelihood that students will successfully engage process 2 (*e.g.* through guided questioning) [22].

Studies [26, 27, 12, 28] described below draw on one or both approaches to address student reasoning and add to the collection of DPTs-based interventions in physics. Note that screening-target question sequences are implemented in each case in a similar fashion to [24].

Gette *et al.* [26] implemented a variety of pedagogical techniques using approach (a) described above designed to address student reasoning inconsistencies on a set of questions related to buoyancy. Their interventions focused on making relevant knowledge more ac-

cessible for students and reducing the appeal of unproductive, intuitive process 1 ideas. In this way, the interventions were more geared towards approach (a) in that they aimed to promote a productive process 1 model as opposed to approach (b) which supports student engagement of process 2 thinking. Two main instructional strategies were implemented. One involved additional explicit instruction on the concept of density which was considered to be more readily available (*i.e.* accessible) in the buoyancy context than the concepts of forces and pressure, which were the focus of original instruction. The second instruction strategy used in the study was to incorporate hands-on and real-world examples. While the density instruction had no impact on performance on the buoyancy target question, the hands-on and real-world applications led to a significant increase in performance. The authors suggest that performance may have increased because the instructional modifications were able to remove the strong intuitive appeal of the common incorrect response. Consequently, the most plausible default model was then often consistent with the correct response. In other words, it seems that the intervention was responsible for the increased generation of correct process 1 ideas.

Another study, conducted by Kryjevskaia *et al.* [27], evaluated student performance on a pair of screening and target questions related to forces and Newton's laws. They developed interventions designed to help students both strengthen relevant content knowledge and support engagement of cognitive reflection skills, in line with approach (b) outlined above. The intervention involved multiple stages including individual and group work. After each stage, students were given the opportunity to revisit their response to the target question. Students were also given a similar post-test question on an exam. The results suggested that every intervention stage improved student performance and, consistent with DPTs, that student success on the post-test was largely dependent on students' conceptual understanding and cognitive reflection skills. No one intervention stage was better than another at improving student performance, and there appeared to be no "quick fix" to helping students overcome persistent intuitive ideas.

In [12], Speirs *et al.* conducted research in the context of various physics tasks to test DPTs-based predictions about student reasoning. They implemented a reasoning chain construction task format for students to answer and explain their reasoning on a target question. In this format, students were provided with reasoning elements (all true statements) and connecting words, which they could assemble into an explanation leading to a conclusion. As part of the experiment, an *analytic intervention element* (AIE) was included in the reasoning chain task for some students. This element provided information refuting an incorrect default model generated by process 1. The authors found that including this element caused more students to abandon the incorrect model compared to a control group that did not receive the AIE. This impact was more dramatic for students who demonstrated successful application of relevant content knowledge on an associated screening question. These results support a DPTs interpretation of student inconsistencies and highlight a way of leveraging dual-process mechanisms to provide a pathway for overriding faulty intuitive ideas. While the AIE intervention was not intended to increase student endurance in cognitive effort, it did work to support process 2 engagement, a primary objective of approach (b) defined above.

In contrast to each of the studies mentioned above, Lindsey *et al.* [28] compared the short-term effectiveness of both approaches (a) and (b) in the context of objects falling at terminal speed in the presence of air resistance. In this experiment, a control group was given an intervention that provided additional practice applying the correct physics reasoning to tasks involving objects falling at terminal speed. The intent was to improve student fluency and therefore, increase the likelihood that the correct reasoning would "come to mind" in a new situation, in line with intervention approach (a). The treatment group was given an intervention that made students aware of possible errors in a common incorrect intuitive idea and guided them in reconciling this intuitive model with correct reasoning, following approach (b) to support students' effective process 2 engagement. All students were then assessed on an analogous post-test question and a near-transfer question testing the same content knowledge in a slightly different way. At the two institutions where the data was

collected, the control and treatment groups were compared across several measures for students demonstrating basic conceptual understanding. At both institutions, the treatment group had larger gains in performance between the original target question and both the post-test question and near-transfer question. Additionally, the treatment group showed better absolute performance on the transfer task at one institution. These results suggest that guiding students in the practice of engaging process 2 is more effective at helping some students succeed on tasks that tend to elicit an unproductive intuitive idea in this context.

These studies collectively demonstrate the application of various screening-target question sequences to disentangle student conceptual understanding and reasoning approaches under the framework of dual-process theories. They also represent an assortment of intervention strategies designed to test the explanatory power of DPTs on student inconsistencies in physics as well as to aid student success on physics questions that often prompt an incorrect process 1 model.

The research described here in Chapters 3 and 4 provides another variety of intervention that takes approach (b) by focusing on three main design principles: the intervention should raise doubt in an unproductive process 1 thought, cue relevant content knowledge, and encourage productive process 2 engagement. The intervention structure involves a fictitious student dialogue and asks students to consider each fictitious student's line of reasoning by applying it to the target question. Then students are asked whether or not they would like to revisit their answer to the target question. After encountering this intervention, many students changed their original target response from an incorrect answer to a correct (large effect size). This work corroborates those outlined above and provides a different strategy for helping students overcome reasoning inconsistencies.

1.7 Dissertation Outline

This dissertation consists of three research articles related to the goals discussed above. Chapter 2 is an article that has been published in *Physical Review Physics Education Re-*

search co-authored with Paula Heron, Chapter 3 is an article in preparation for submission to the same journal also co-authored with Paula Heron, and Chapter 4 is an article describing research that extends the work done in Chapters 2 and 3.

Chapter 2 focuses on the first goal, to establish a model for validating sets of screening and target questions and to evaluate the sufficiency of a dual-process theory interpretation of related data. Over 1,200 undergraduate students in introductory calculus-based waves and optics courses at the University of Washington were given screening and target questions in the context of a pulse on a spring adapted from a study by Kryjevskaia *et al.* [24]. Student response patterns were analyzed alongside their scores on the cognitive reflection test (CRT), a three-item measure of one's propensity to mediate intuitive ideas with cognitive reflection [21]. The purpose of this analysis was to determine if two assumptions held: (1) that the screening question adequately screens out students lacking relevant conceptual understanding and (2) that correctly responding to the target question is largely a matter of cognitive reflection on incorrect intuitive ideas for students who have adequate conceptual understanding. The data generally supported a DPTs perspective of student reasoning in this context. The screening and target questions also appeared to be functioning as intended to distinguish between students' conceptual understanding and reasoning approaches.

Chapter 3 extends the research regarding the pulse question sequence by focusing on an intervention, which addresses the second main goal of this work. This paper outlines the intervention design grounded in dual-process theories, details its iterative development, and provides analyses of data from almost 630 students. The results showed that the intervention was effective at raising doubt in students' unproductive ideas generated by process 1 and supported productive process 2 engagement by encouraging students to consider alternative lines of reasoning. The intervention equally benefited students regardless of their natural tendency towards cognitive reflection. This chapter demonstrates the short-term benefits of a micro-intervention strategy in a single context, which can be applied to other contexts.

Chapter 4 is a research article that addresses both goals (1) and (2). This study used a

question sequence similar to that outlined in Chapters 2 and 3 but in a different context involving boxes on rough surfaces. This new sequence tested the screening-target sequence validation methods introduced in Chapter 2 and evaluated the effectiveness of intervention strategies developed in Chapter 3 but applied in the box-friction context. Results supported the two DPTs-based assumptions regarding the screening and target questions (see assumptions above). The data also indicated that the intervention was beneficial for students. Students who incorrectly responded to the target question were more likely to indicate a desire to revisit their response after answering questions that were intended to support analytic thinking compared to students who did not see these analytic support questions. Additionally, students improved performance on the target question after encountering the intervention. Both of these results were more pronounced for students who correctly answered the analytic support questions, as expected, and students seemed to benefit regardless of their propensity for cognitive reflection. Overall, the chapter demonstrates that the design and evaluation methods developed in the pulse context were able to carry over into the forces context.

Chapter 2

DISTINGUISHING BETWEEN STUDENTS' CONCEPTUAL UNDERSTANDING AND REASONING APPROACHES: AN APPLICATION OF DUAL-PROCESS THEORIES

Kristin Kellar and Paula Heron

Department of Physics, University of Washington, Seattle, Washington 98195

2.1 Abstract

Dual-process theories of reasoning suggest that humans reason using two processes often referred to as process 1 (heuristic) and process 2 (analytic). When presented with a situation requiring any sort of reasoning or decision-making, process 1 automatically engages and generates an initial mental model to address the situation. Process 2 may or may not be engaged to assess the initial model as a plausible solution. In a study by Kryjevskaia *et al.*, a "screening" question regarding a pulse on a spring aimed to identify students with relevant content knowledge who nevertheless seemed to rely on process 1 when answering a subsequent "target" question. The study was offered as evidence that dual-process theories can explain some discrepancies in student responses to related questions. The study described here assesses the same pair of questions for their ability to distinguish between incorrect answers that stem from inadequate conceptual understanding and those that stem from reasoning approaches. We use Frederick's cognitive reflection test as part of this analysis. Our results largely support a dual-process-theories perspective of student reasoning.

2.2 Introduction

Physics education research has documented student conceptual difficulties in various contexts. While inadequate conceptual understanding is the cause of some student errors in physics, there is evidence to suggest that even students who have developed considerable conceptual understanding do not necessarily answer relevant questions correctly. For example, students who exhibit correct conceptual understanding on a particular task may fail to utilize the same knowledge and skills on a similar task, even when those tasks are presented consecutively [24, 29, 27, 22].

It has been hypothesized that such responses can be explained by dual-process theories (DPTs). These theories describe human reasoning as consisting of two processes: an automatic/heuristic process (process 1) and a deliberate/analytic process (process 2) [19]. This paper presents research that aims to gain insight into introductory physics students' reasoning processes and validate a set of questions designed to disentangle the effects of reasoning approaches and conceptual understanding.

We used a question sequence centered around pulse propagation designed to identify students who had appropriate content knowledge yet made reasoning errors. First a "screening" question was presented to students designed to screen for conceptual understanding. A "target" question followed that necessitates the same content knowledge and skills as the screening question but tends to cue a misleading intuitive idea. The authors of the original pair of pulse propagation tasks [24] concluded that students who demonstrated a reasonable understanding of the physics content on the screening question but failed to correctly answer the target question were relying on process 1, rather than searching for alternative models using process 2. We have examined student responses to this pair of questions, as well as several related ones, and also designed an intervention to help students ultimately arrive at the correct conclusion on the target question. In this paper we focus on validation of the screening-target question pair, specifically examining the conclusions drawn by Kryjevskaia

et al. [24] about student responses to both the screening and target questions. A different paper [30] provides a thorough examination of the intervention.

This paper starts by outlining the underpinnings of dual-process theories and how prior research informed our investigation. We then describe the question sequences used in the study before presenting results. These results are discussed in the context of theory and other research, and suggestions for further research are provided.

2.3 Theoretical Framework

Demonstrating adequate content knowledge on one question but failing to demonstrate that same knowledge on a related question is characteristic of dual-process theories of reasoning. As noted above, DPTs center on a common description of the reasoning process: In one's thinking, reasoning, or decision-making, there are two main processes. process 1 is automatic, and intuitive, and often fast while process 2 is deliberate, rule-based, and often slower than process 1 [19]. There are many different "flavors" of DPTs that have various applications and emphasize the interaction between the two processes in different ways. Our collaboration has been operating under the DPTs-informed framework published in [22], which is presented in Figure 1 and described in detail below.

When presented with a problem or situation that warrants a decision, process 1 is activated automatically and creates an initial model for addressing the problem. The initial model that is created draws from prior experience and contextual cues. It represents a sort of "first impression" or "gut feeling" about the problem at hand. We will refer to this initial reasoning response as "intuition" in the sense that it involves knowledge that comes automatically to mind after being presented with a task. Intuition can involve formal physics knowledge or informal ideas. Either way, intuition cannot be turned off. When engaged, process 2 works to evaluate, analyze, or justify the initial model. If, upon evaluation, the model is deemed satisfactory, it becomes the final answer or conclusion. If during this analysis, a problem is identified, a new model might be created, and the cycle repeats until a conclusion is reached.

When addressed with a task, there are different ways one can engage process 1 and process 2 to ultimately reach a conclusion. In particular, there are two main reasoning pathways referred to as *the path of cognitive frugality* and *the path of sustained effort* [22]. In each path, the reasoner may encounter various pitfalls or hazards that may lead to incorrect conclusions. A reasoning pathway diagram, including possible hazards is shown in Fig. 2.1. Each reasoning pathway begins when a task is presented and process 1 automatically generates a provisional model based on cued knowledge. During this part of the reasoning process, one possible pitfall, hazard A, presents itself when process 1 generates an unproductive intuitive model, (unproductive in the sense that it is incorrect or leads directly to an incorrect answer). This hazard may occur when one lacks relevant "mindware," defined as stored content knowledge, rules, procedures, and strategies necessary to successfully complete the task at hand [20]. Even when mindware is adequate, one may encounter hazard A if a task tends to cue irrelevant or incomplete knowledge. Regardless of accuracy, once a provisional model is created, the reasoner may either reflect or use the model as a final decision. It is at this junction where the path of cognitive frugality and the path of sustained effort part ways.

In the path of cognitive frugality, at the junction between reflection or using an initial model as the final decision, process 2 is not engaged; the reasoning process is governed by process 1. In everyday situations, an initial model is often acceptable and so process 2 may not be needed. Only when deemed necessary does the human reasoning process put forth the extra effort involved in utilizing process 2. In this way, we tend to use processing mechanisms that result in low computational expense. Following [22] we refer to the propensity for low computational expense in such a case as "cognitive frugality," hence the moniker "path of cognitive frugality" for the reasoning pathway described above. While this tendency is computationally efficient, it also leads to results that are incorrect if process 1 generates a poor initial model [20]. This may be especially problematic in physics where many concepts are counterintuitive and reflection and subsequent analysis may be needed, at least for novices. (Experts may generate correct initial responses to questions on elementary topics, allowing

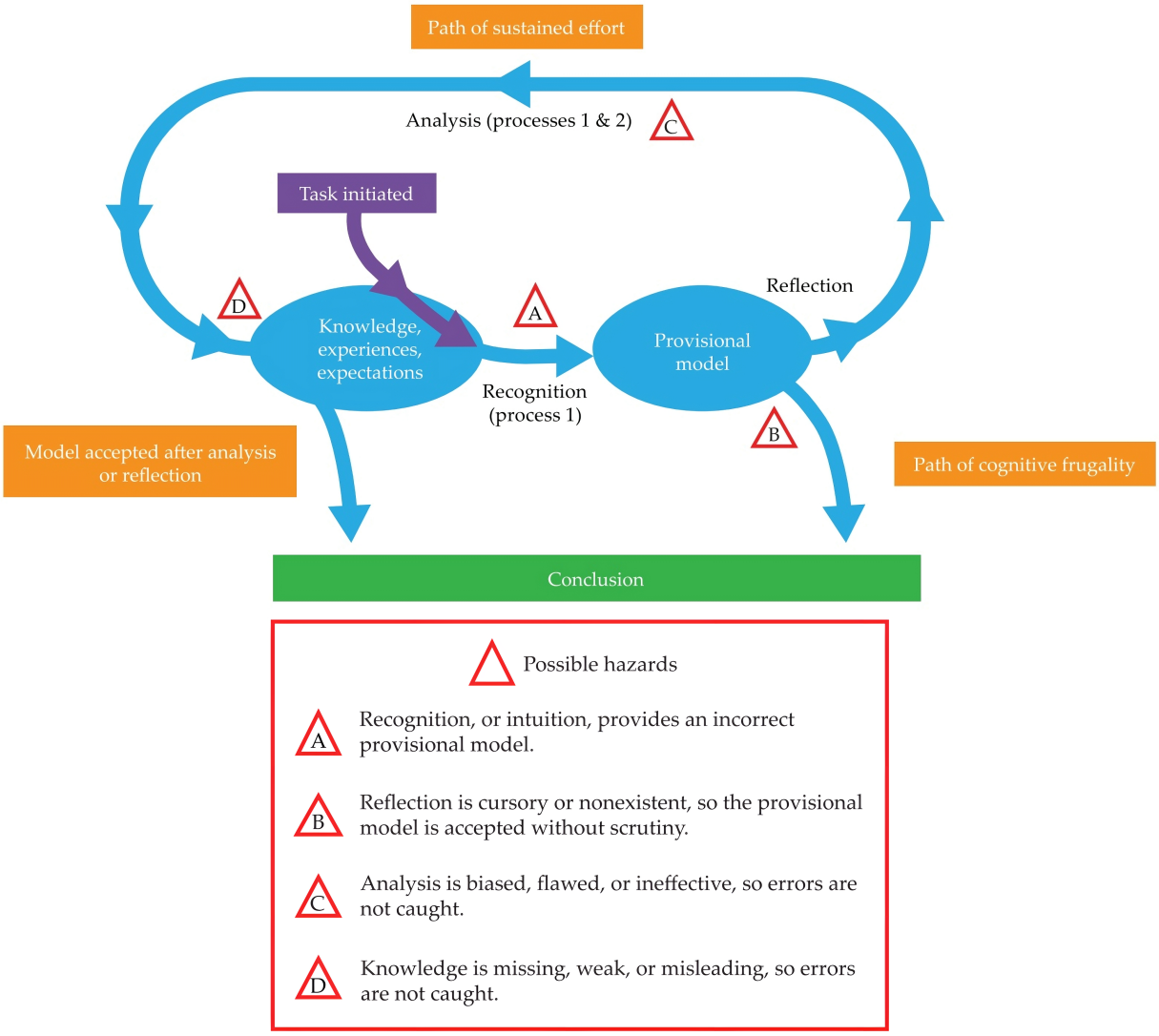


Figure 2.1: Diagram from [22].

them to avoid cognitive effort and efficiently reach correct responses [22].)

For students who fall prey to hazard A by developing an unproductive intuitive model, taking the path of cognitive frugality by accepting their provisional model without scrutiny is considered hazard B. There are two main causes for reasoners to experience hazard B, inadequate mindware and detection failure. Without the relevant mindware, one does not recognize the need to engage the analytic thinking process, and so a misleading intuition becomes the final conclusion. In detection failure, even while capable due to adequate mindware, one does not detect the necessity of using process 2 to analyze one's initial model [20].

The path of sustained effort leads the reasoner to analysis. There are two main classes of process 2 engagement during this analysis phase. First, the natural tendency is to look for evidence to support the initial model as opposed to utilizing a falsifying strategy of hypothesis testing [23]. This is known as rationalization [31]. The second class of analytic thinking is known as cognitive decoupling, which involves processing that inhibits and overrides an intuitive response [31].

Yet again, there are reasoning hazards that can occur during the path of sustained effort, which lead to what is referred to by Stanovich [20] as sustained override failure. Sustained override failure

occurs when process 2 is activated, but not sustained, and so an initial unproductive model is not replaced with the correct one. One reason is hazard C, biased reasoning in the form of rationalization or confirmation bias, which merely validates the initial model. Additionally, even if process 2 tests and challenges the provisional model, hazard D, characterized by weak mindware (as discussed earlier) may prevent the reasoner from being able to replace an incorrect model with a sound one.

If during the reasoning process, one manages to avoid hazard B and engage in the path of sustained effort in such a way as to produce a new provisional model, the cycle can repeat.

Since intuition cannot be turned off, and is not always correct, it is useful to be able to

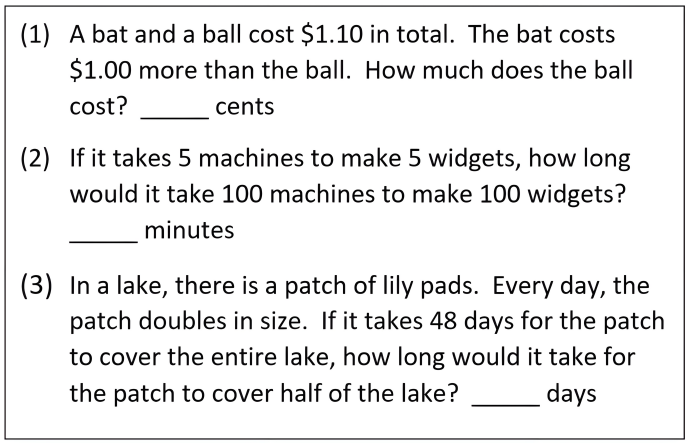


Figure 2.2: The cognitive reflection test, developed by Frederick [21], purports to measure one's ability to override an incorrect intuitive thought and replace it with the correct answer.

effectively engage process 2 to reflect on an initial thought and ultimately produce the correct answer. The ability or disposition to override the response that first comes to mind is called "cognitive reflection." The propensity for cognitive reflection is often measured by a three-item test called the cognitive reflection test (CRT), seen in Figure 2.2 [21]. Each question on the test requires minimal mindware and tends to trigger an incorrect initial model. In order to answer correctly, one must engage the analytic thinking process and override the first idea generated by the heuristic process. Therefore, a higher score of 2 or 3 is interpreted as indicating that an individual has engaged in greater cognitive reflection while a lower score of 0 or 1 suggests lower cognitive reflection[1]

2.4 Background and Motivation

Kryjevskaia *et al.* [24] developed two sets of physics questions with the intent of disentangling students' conceptual understanding and reasoning approaches: one on pulse propagation and one on capacitors. First, students are asked one or more "screening" questions that require them to apply mindware in a specific area of physics. Then, students are asked a

[1]A paper discussing the CRT in further detail is in progress. Additionally, further analysis is being conducted using the seven-item CRT.

"target" question, which appears, to an expert, to require the same mindware as the screening question. However, the target question typically involves a context or feature that tends to elicit an unproductive initial model. These types of contexts or features are referred to as "salient distracting features" [14, 32, 33].

If a student incorrectly responds to the screening question(s), it is an indication that they may lack the mindware required by the target question. On the other hand, if a student answers the screening question(s) correctly, they probably have adequate mindware. Of course, with a single question it is possible that some students are misclassified, but on average it is assumed that most students who answer correctly do have adequate mindware. The ability to access adequate mindware on the screening question(s) presupposes the same is possible for the target question, given the design of the screening-target question pairing.

Using two sets of screening and target questions, Kryjevskaia *et al.* [24] showed that some students who demonstrated conceptual understanding on a screening question failed to do so on a subsequent target question. The salient distracting features of the target question were clearly implicated in incorrect answers given in response to the target question. These observations were explained in terms of dual-process theories, namely that many students relied on process 1 for the target question, rather than effectively engaging process 2. This interpretation relies on two important assumptions: (1) the screening question accurately classifies students as having, or not having, adequate conceptual understanding to answer the target question correctly; and (2) among content-proficient students (those who answer the screening question correctly), responding to the target question is largely an issue of cognitive reflection on misleading process 1 answers. That is, students who reflect may go on to answer correctly, while students who do not, answer incorrectly.

The first assumption implies that the screening and target questions require the same concepts to answer, or at least that those who have the conceptual understanding necessary to answer one, have the conceptual understanding necessary to answer the other. We argue that if this is correct, then: (i) students who answer the screening question incorrectly

should not be able to answer the target question correctly (*i.e.*, a correct target response requires understanding of the concepts tested by the screening question); and (ii) students who answer the screening question correctly should be able to answer related questions that also lack salient distracting features.

We also argue that if assumption (2) above is correct, then: (iii) content-proficient students who exhibit a higher propensity for cognitive reflection should be more likely to answer the target question correctly; and (iv) interventions designed to support students in reflecting on their thinking, without promoting further conceptual development, should increase the number of correct responses to the target question.

In this paper we examine these assumptions in greater detail. In addition to its relevance to the claims made by Kryjevskaia *et al.* [24], we believe our analysis can serve as a model for assessing the sufficiency of DPTs interpretations, and for identifying valid screening-target sequences.

2.5 Investigation

This study focuses on student responses to a screening-target question pair from [24], a set of related questions that we designed, and the cognitive reflection test [21]. The screening-target pair was part of more extensive question sequences that also included a set of intervention questions that were designed to encourage students to effectively engage in analytic thinking. The intervention questions are briefly touched on here, but a deeper discussion can be found in [30].

2.5.1 Context

This study was conducted in introductory calculus-based physics courses at the University of Washington (UW) during six quarters starting in Autumn of 2020 and ending in Spring of 2022 (N = 1,223). The courses covered material on waves and optics and included lecture, lab, and tutorial components.

The question sequences used in this study were added to online quizzes given after relevant lecture instruction, but before small-enrollment tutorial sections on the topic emphasized on the quiz. These quizzes, called "tutorial pretests," were available online typically for two and a half days over a weekend. Students were typically limited to 15 minutes to complete each one, which was intended to be sufficient for completion without the need to rush. When our question sequences were added to the beginning of a pretest, the time limit was extended to 30 minutes. Students earned a completion grade for attempting each quiz regardless of the correctness of their answers to the questions, and these completion grades comprised a very small percentage of each student's total course grade. As the online quizzes were already part of the course structure, including the question sequences in this way allowed for minimal classroom disruption.

The CRT was separately administered as part of these online quizzes and repeated in each course every quarter so that some students took the test more than once. In such cases, their first CRT score was used for analysis.

Only results from students who consented to participate in the study (or did not opt out of participation) were included in our analysis.

2.5.2 Question sequence overview

Over the course of several quarters, we refined our question sequence. In some cases, minor modifications were made to make the questions clearer. In other cases, we introduced additional questions to probe student conceptual understanding. In total, four versions were developed.

2.5.3 Question sequence components

While this study focuses on student responses to only some questions, it is worth noting that these questions were part of a larger sequence. The structure of each sequence is shown in Fig. 2.3. In each, screening and target questions were followed by a separate page containing

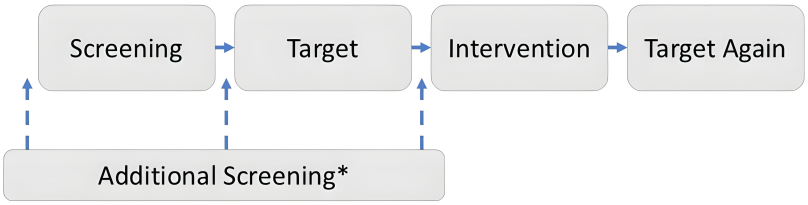


Figure 2.3: The main components of our entire question sequence. *Sometimes all or part of the additional questions inserted after the target question were part of an intervention, but this is not important for the purposes of this paper.

intervention questions, ultimately ending in a subsequent page containing a repeat of the target question. Once students moved from one page of the quiz to the next, they were not allowed to go back. In some versions, additional screening questions were added surrounding the screening and target questions in various configurations.

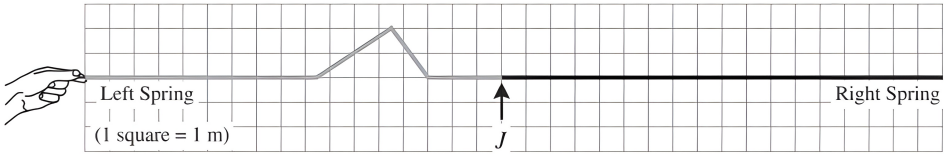
2.5.4 Questions administered

This section introduces some of the questions asked of students that will be relevant for this paper. In particular, the screening and target questions are discussed in detail. In addition, we also discuss a set of questions we refer to as the "bug questions," which were relevant for assessing the effectiveness of the screening question. While the intervention questions are not included in this section, a brief discussion can be found at the end of Sec. 2.6. For an extensive discussion of these questions, see Ref. [30].

Screening and target questions

The pulse screening and target questions developed by Kryjevskaia *et al.* [24] were used in all iterations of our question sequence. On all versions, we divided the target question into two separate questions, one multiple choice and one short answer for easy analysis. On later iterations, these questions were nearly identical except for modified wording for clarity (see Fig. 2.4).

In experiment 1, two different springs are connected at a junction point *J*. A student generates a pulse on the left spring shown below. It takes the student's hand a time Δt_0 to quickly move the end of the spring back and forth in order to generate the pulse. In experiment 1, the propagation speed of a pulse on the left spring is 1.5 times that on the right spring (v_L=1.5v_R).



Experiment 2 is nearly identical to experiment 1 except for a ***single change.*** As a result of this change, the width of the generated pulse (in the left spring) is doubled. The tension in the spring on the left and the time it takes for the student's hand to move to create the pulse is the same in both experiments. The spring on the right is unchanged.

1. Determine the change that has been made in experiment 2. Explain.

2. Is the width of the ***transmitted pulse*** on the right spring in ***experiment 2*** *greater than, less than,* or *equal to* the width of the ***transmitted pulse*** on the right spring in ***experiment 1***?

3. Explain your response to the previous question.

Figure 2.4: Screening and target questions adapted from [24] as used on versions 2 and 4 of our question sequence. (The screening and target questions used on version 1 were identical to those in [24] before minor wording adjustments were made, and version 3 featured two insects as part of the setup that did not affect the pulse (see Fig. A.1 in Appendix A).)

The scenario students were asked about involved left and right springs connected at a junction point J. In experiment 1, a student's hand moved up and down in a time Δt_0 to generate a pulse on the left spring. In experiment 2, a change was made such that the width of the generated pulse was doubled. Students were told that the time for the hand to move and the tension in the springs had not changed. The screening question asked students to determine the change that was made in experiment 2. The target question asked students to compare the width of the pulse in the right spring in each experiment.

To answer the screening question correctly, students needed to understand what variables affect the width of a pulse. One relevant relationship is that pulse width is directly proportional to both the time to generate the pulse and the pulse speed. Pulse speed itself is related to the medium through which the pulse travels. Formally, $v = \sqrt{\frac{F_T}{\mu}}$, where v is the pulse speed, F_T is the tension force, and μ is the linear mass density of the medium. It follows that in order to double the pulse width while holding the time to generate the pulse constant, the wave speed must have doubled. Using the relationship between speed and the medium, students can reason that the linear mass density must have been decreased by a factor of four, since the problem specified that tension remained the same.

The correct response to the target question required the same mindware needed on the screening question because in this case, students should have recognized that given there was no change in the tension or the right spring between experiments, the speed of the pulse when propagating through the right spring would not change. As the time to generate the pulse remained constant, the width of the pulse would be equal in both experiments.

Additional questions

It occurred to us that even students who answered the screening question correctly might not be aware that the time for the pulse to pass the junction would be the same in both experiments. Therefore, on each version of the question sequence, we included a set of questions that encouraged students to think about pulse transmission. They were originally

used as intervention questions and were later repurposed and modified to act as additional screening questions. We used students' various unproductive lines of thought, as documented by Kryjevskaia *et al.* [24] from student responses to the target question, to inform this process. Most students who answered the target question incorrectly in that study explained their answer using the idea that the width of the generated pulse had a direct effect on the width of the transmitted pulse. Since the width of the generated pulse changed in experiment 2, they assumed the width of the transmitted pulse changed in a similar fashion, either because they reasoned that a faster incident pulse would result in a faster transmitted pulse, or that the time for the transmitted pulse to form increased due to the increased width of the incident pulse. These arguments are counter to the stated facts that the generation time and the medium in which the transmitted pulse traveled did not change. With this in mind, we developed sets of "bug questions" (used on versions 1 through 3 of the question sequence) that emphasize the time for the junction to move up and down as the pulse passed in each experiment of the screening-target question setup.

When used as additional screening questions, the bug questions provided us with more information about the mindware students could demonstrate. The bug questions used in versions 1 through 3 were very similar, mainly differing in wording and location in the sequence. Administering the bug questions before and after the target question on versions 3A and 3B allowed us to test if the order affected student performance on the bug questions, which it did not. We will outline each set of bug questions below.

The setup for the bug questions was identical to that used for the screening and target questions except for the addition of a ladybug resting on the student's hand and a bumblebee resting on the junction J. (See Appendix A for modified questions used in versions 3A and 3B.) The text indicated that these insects were lightweight and did not affect the pulse in any way. On version 1, students were asked to compare the time for the bumblebee to move up and down to that of the ladybug in experiments 1 and 2 (before and after a change was made to double the width of the generated pulse). In either case, the time for the bumblebee

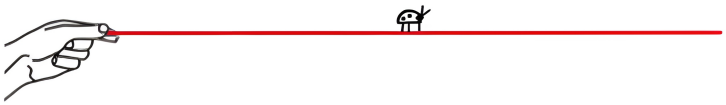
to move up and down would be equal to that for the ladybug since the time for the vertical motion depends on the source generating the pulse and is independent of the medium.

In an effort to gain a more comprehensive picture of students' conceptual understanding of pulse propagation, on version 4 of the question sequence, the bug questions were changed to address more aspects of pulse motion. The new version 4 bug questions encourage students to think about the transverse motion of the medium, and also the motion of the pulse propagating parallel to the medium.

The new bug questions on version 4 (see Fig. 2.5) asked about both parallel and transverse motion for a pulse traveling along a rope using a new scenario. In this case, in experiment A a student generated a pulse on a rope that had a ladybug of negligible mass resting on its middle. In experiment B, the student moved her hand up and down in half the time compared to experiment A. Question 1 asked students to compare the time for the leading edge of the pulse to reach the bug in each experiment, prompting students to think about the motion of the pulse parallel to the medium. The correct answer requires recognizing that since the medium was unchanged from experiment A to experiment B, the pulse speed was unchanged, and therefore the time for the leading edge to reach the bug was the same in each experiment. Question 2, which focused on transverse motion, asked students to compare the time for the bug to move up and down in each experiment. Since the time for the hand to move up and down to generate the pulse was halved in experiment B, the time for any point along the medium to move up and down would be halved as well, including the point at which the bug was located. The correct answer is therefore that the time is less in experiment B than experiment A. Question 3 prompted students to explain their responses.

2.5.5 Summary of question sequence iterations

A summary of the screening and target questions for each version of the question sequence is listed in Table 2.1. See above for details about the screening, target, and bug questions.



In experiment A, a student moves her hand up and down in a time Δt_0 to generate a pulse on the rope shown. A bug of negligible mass rests in the middle of the rope. In experiment B, the student moves her hand up and down with the same amplitude as before, but in half the time: ½ Δt_0.

1. Is the time for the leading edge of the pulse to reach the bug in experiment B *greater than, less than,* or *equal to* what it was in experiment A?
2. Is the time for the bug to move up and then back down again in experiment B *greater than less than,* or *equal to* what it was in experiment A?
3. Explain your responses to questions 1 and 2.

Figure 2.5: New bug questions used on the version 4 question sequence. These questions focused on both parallel and transverse motion during pulse propagation. Questions 1 and 2 were multiple choice, and question 3 was in short answer format.

Table 2.1: Summary of question sequence iterative development process showing question types in the order given for each version. Elements that were changed from preceding versions are in bold. *Asterisks indicate versions during which instruction was virtual due to the COVID-19 pandemic.

Version	Additional screening	Screening	Additional screening	Target	Additional screening
1*		Original from [24]		Original from [24]	Bug Q's
2*		**Minimally Modified**		**Minimally Modified**	**Modified Bug Q's**
3A*		**Minimally Modified (Bugs in Diagram)**	**Modified Bug Q's**	**Minimally Modified (Bugs in Diagram)**	
3B*		**Minimally Modified (Bugs in Diagram)**		**Minimally Modified (Bugs in Diagram)**	**Modified Bug Q's**
4	**New Bug Q's**	Minimally Modified (same as 2)		Minimally Modified (same as 2)	

2.6 Results and Discussion

In this section we present an overview of results from our administration of the screening and target questions, then discuss specific results related to our main goals of testing two assumptions about the screening and target questions: (1) that the screening question accurately classifies students as having, or not having, adequate conceptual understanding to answer the target question correctly; and (2) that among content-proficient students (those who answer the screening question correctly), responding to the target question is largely an issue of cognitive reflection on misleading process 1 answers.

2.6.1 Overview of results

We classified responses to the screening question as "correct" if they identified a valid factor that affects pulse speed, regardless of whether the relationship between that factor (*e.g.,* linear mass density) and pulse speed was accurate. Therefore, we considered as "correct" any response that indicated that the only change made to the experiment involved the medium or its properties (weight, spring constant, mass density, *etc.*). We refer to students who gave these responses as " content proficient." Responses to the target question were considered correct based on the selection of the multiple-choice option "equal to" regardless of accompanying explanations. The data analyzed in [24] were obtained from course exams on which students received credit for their explanations. Thus, the authors took into account the accuracy of student explanations in determining correctness. However, our data were obtained from quizzes that were graded only for completion. Many student explanations were so brief that we were unable to follow the arguments sufficiently well to determine the accuracy of their reasoning. In any case, we would need to use the same criteria as the authors of [24] to make any direct comparisons meaningful, criteria that were not included in that paper.

Since we were primarily interested in students who exhibited mindware but answered the target question incorrectly, we will define two groups of content-proficient students. "All-

Table 2.2: Definitions

Term	Definition
Content-proficient students	Students who answered the screening question correctly
All-correct students	Content-proficient students who answered the target question correctly
Reasoning-error students	Content-proficient students who answered the target question incorrectly

Table 2.3: Student performance on the pulse screening and target questions given on all versions of the question sequence compared to student performance in the prior study done by Kryjevskaia *et al.* [24]. Note that when the percentage of content-proficient students was calculated for each version individually, the range was 41% to 69% (which is in line with random variation to be expected for this type of data [34]). Likewise, the range of percentages for the next two columns was 45% to 58% and 42% to 55% respectively.

	Percent of total who were content proficient	Percent of content-proficient students who were all correct	Percent of content-proficient students who made reasoning errors
[24] (N=169)	75%	69%	31%
All versions (N=1,223)	51%	53%	47%

correct" students are those who answered both the screening and target questions correctly. "Reasoning-error" students are those who answered the screening question correctly but went on to answer the target question incorrectly. Table 2.2 contains these definitions for easy reference.

Since the screening and target questions were essentially identical across all versions of our question sequence, and we saw no significant differences in the fraction of content-proficient students who correctly answered the target question, we pooled the data to compare our results with those reported by Kryjevskaia *et al.* [24], as shown in Table 2.3.

It is evident that there is a lower percentage of students who answered the screening question correctly as compared to the prior study [24], especially given the more stringent conditions for correctness in that study. It also seems that the proportion of UW students who were able to answer the screening question correctly but went on the answer the target question

incorrectly, is greater than in [24]. The differences may be due to the fact that our question sequence was administered after relevant lecture instruction but before tutorial instruction, while the questions in the previous study were given during a course exam after students had more experience with the content. Regardless, our results are largely consistent in that a significant fraction of students who demonstrated mindware on the screening question failed to answer the target question correctly.

2.6.2 Validation of the screening question (testing assumption 1)

As mentioned in Sec. 2.4 above, if the screening question operates as intended (*i.e.* accurately classifies students as having, or not having, adequate conceptual understanding to answer the target question correctly), then (i) students who answer it incorrectly should not be able to answer the target question correctly, and (ii) students who answer it correctly should also be able to answer related questions that also lack salient distracting features.

Students' target performance predicts screening performance

Our hypothesis was that students who answer the target question correctly have the necessary mindware in addition to cognitive reflection skills, whereas the screening question requires mindware but minimal cognitive reflection. If the screening question requires the same mindware as the target, then students who answer the target question correctly due to application of relevant mindware should also answer the screening question correctly. In this way, correctly answering the target question would imply correctly answering the screening question.

We calculated a risk ratio to determine if accuracy on the target question affected the likelihood of success on the screening question.

Table 2.4: Contingency table for student performance on screening and target questions for all versions.

	Screening incorrect	Screening correct	Total
Target incorrect	498	290	788
Target correct	102	333	435
Total	600	623	1223

$$\begin{aligned}\text{Risk Ratio} &= \frac{\text{Screening correct \& target correct/Target correct}}{\text{Screening correct \& target incorrect/Target incorrect}} \\ &= \frac{333/435}{290/788} = \frac{0.77}{0.37} = 2.08.\end{aligned}$$

As anticipated, answering the target question correctly increased the likelihood of answering the screening question correctly. In fact, students who performed well on the target question were 2.08 times more likely to answer the screening question correctly than those who failed to answer the target question correctly.

A more comprehensive look at student response patterns on the screening and target questions revealed a hierarchy consistent with DPTs, namely that success on the target question necessarily implies accuracy on the screening question. We found a medium-level correlation between performance on the screening and target questions to this effect by conducting a chi square test on the data show in Table 2.4 [$p < 2.2 \times 10^{-16}$, Cramér's V effect size = 0.38 (medium)]. The contingency table presented in Table 2.4 is also represented as a Sankey diagram in Fig. 2.6.

We also followed the method outlined in [35] to investigate a possible hierarchy in student responses. The anticipated hierarchy, that correctly answering the target question necessarily implies correctly answering the screening question, is supported by the fact that 77% of students who answered the target question correctly also answered the screening ques-

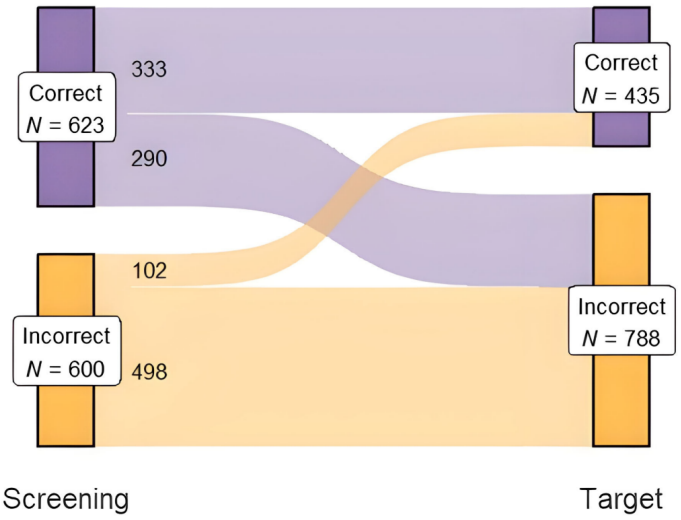


Figure 2.6: Sankey diagram showing pathways of answering patterns on screening and target questions for all versions.

tion correctly. The logically equivalent statement that incorrectly answering the screening question implies incorrectly answering the target question is also supported since 83% of students who incorrectly answered the screening question also answered the target question incorrectly. Consistent with this hierarchy, we see that correctly answering the screening question does not necessarily imply correctly answering the target question given that 290 content-proficient students failed to arrive at the correct conclusion on the target question. Ideally, there should be no students who correctly answered the target question without demonstrating adequate content knowledge by successfully answering the screening question. Our data shows that there were 102 students who fell into this category. We don't have a clear explanation for this, but we suspect it could be due in part to random guessing.

The results largely support the idea that the screening and target question require the same content knowledge, or at least that the content knowledge required to answer the target question, implies the presence of the content knowledge required to answer the screening question.

Table 2.5: Contingency table for student performance on screening and bug questions for all versions.

	Bugs incorrect	Bugs correct	Total
Screening incorrect	301	299	600
Screening correct	205	418	623
Total	506	717	1223

Content-proficient students outperform noncontent-proficient students on additional screening questions

To gain more insight into the ability of the screening question to identify students with mindware, we used another method of analysis. As stated previously, in an ideal situation, students who answer the screening question correctly are equipped with all the mindware necessary to answer the target question correctly. In particular, we wanted to assess whether students were being inaccurately classified as content proficient. (Given that our primary interest was in the reasons that content-proficient students would answer the target question incorrectly, it was important that this pool not contain students whose incorrect answers stemmed from lack of conceptual understanding. We were less concerned if students were being inaccurately classified as not content proficient.) Since the bug questions and the original screening and target questions involved closely related mindware, one would expect that the students who answered the screening question correctly would perform better on the bug questions than those who answered the screening question incorrectly.

The contingency table in Table 2.5 shows the results for all versions combined. Performance on the bug questions is evidently higher overall than on the original screening question, but a chi square test indicates that content-proficient students tended to perform better than noncontent-proficient students [$p = 8.94 \times 10^{-10}$, Cramér's V effect size $= 0.18$ (small)].

The table also shows that the success rate on the bug questions for noncontent-proficient students is approximately 50%. We note that bug and screening questions test *related*, but

not *identical* content knowledge. Our main goal in asking the former was to determine whether the screening question could be used to *screen out* students lacking the content knowledge necessary to answer the target question. If the screening question were perfect, there would be no noncontent-proficient students in the 'content proficient' pool (no false positives), and no content-proficient students in the 'noncontent proficient' pool (no false negatives). In such a case, every student who answered the screening question correctly would also answer the bug questions correctly. Evidently this is not the case, but the actual relationship is strong enough that we feel confident that using the screening question alone minimizes the number of false positives.

At the same time, it appears that incorrectness on the screening question does not similarly imply incorrectness on the bug questions, suggesting there might be a significant number of false negatives. Alternatively correct scores could result from guessing. However, "correctness" requires getting both questions right, which seems unlikely to result from guessing. We can't really distinguish these possibilities, but again, our main analysis depends on the lack of false positives.

2.6.3 Role of cognitive reflection (testing assumption 2)

As mentioned earlier, if content-proficient students who answer the target question incorrectly do so because they don't adequately reflect on misleading process 1 answers then (iii) content-proficient students who exhibit a higher propensity for cognitive reflection should be more likely to answer the target question correctly; and (iv) interventions designed to support students in reflecting on their thinking, without promoting further conceptual development, should increase the number of correct responses to the target question.

CRT scores are linked to performance

Using a DPTs lens, a correct response to the target question is largely dependent on cognitive reflection on misleading process 1 answers. If this is the case, we would expect that content-

Table 2.6: Contingency table of low and high CRT scores for all-correct and reasoning-error students for all versions.

	Low CRT (0-1)	High CRT (2-3)	Total
All correct	69	264	333
Reasoning error	96	194	290
Total	165	458	623

proficient students who responded to the target question correctly generally had a higher propensity to mediate their intuitive thoughts with analytic reasoning. Measurably, we predicted that content-proficient students who answered the target question correctly would have higher scores on the CRT than content-proficient students who answered it incorrectly. Similarly, content-proficient students with high CRT scores would be more likely than those with low CRT scores to answer the target question correctly. We would also predict that for content-proficient students, every one-point increase in CRT score would correspond to a higher probability of answering the target question correctly. Below we discuss our tests of these predictions.

Content-proficient students' CRT scores are associated with target performance We predicted that all-correct students would have higher scores on the CRT than reasoning-error students. We grouped students according to whether their CRT scores were high (2-3) or low (0-1). On all versions, a greater proportion of all-correct students earned high CRT scores than reasoning-error students. We conducted chi square tests of independence for each version individually. Version 3 showed no difference in CRT scores between all-correct and reasoning-error students. All other versions showed small or moderate effect sizes. For pooled data from all versions (see Table 2.6), a chi square test produced a small effect size [$p = 0.0005$, Cramér's V effect size $= 0.14$ (small)].

There are several reasons that there might not be large differences in CRT scores between all-correct and reasoning-error students. One is that the screening question itself might

Table 2.7: Contingency table of low and high CRT scores for noncontent-proficient and content-proficient students for all versions.

	Low CRT (0-1)	High CRT (2-3)	Total
Screening incorrect	217	383	600
Screening correct	165	458	623
Total	382	841	1223

require some cognitive reflection. In this case, it is likely that content-proficient students would generally have higher CRT scores so there would be less of a difference in CRT scores between students within this group who answered the target question differently. In order to test this possibility, we conducted a chi square analysis to compare the proportions of content-proficient and noncontent-proficient students earning high and low CRT scores (see Table 2.7). If the screening question required the application of cognitive reflection skills in addition to relevant mindware, we would expect that students who answered the screening question correctly would have higher CRT scores than students who did not answer the screening question correctly.

For all versions combined, the results were significant [$p = 0.0003$, Cramér's V effect size = 0.10 (small)]. On each question sequence version analyzed, a higher proportion of students who answered the screening question correctly (content-proficient students) scored high (2-3) on the CRT compared to those who were not content proficient. (As with the previous analysis related to the CRT, when analyzed individually version 3 was the only one that did not produce significant results.) It appears that the screening question required some cognitive reflection, which could be making the association between CRT score and target performance appear less strong.

To gain a better understanding of the association between content-proficient students' CRT scores and their performance on the target question, we also calculated a risk ratio for all versions combined. The goal of this test was to compare content-proficient students'

likelihood of answering the target question correctly (belonging to the all-correct group) depending on their CRT scores using the data shown in Table 2.6. Specifically, we observed the following ratio:

$$\begin{aligned}\text{Risk Ratio} &= \frac{\text{All correct high CRT/Content proficient high CRT}}{\text{All correct low CRT/Content proficient low CRT}} \\ &= \frac{264/458}{69/165} = \frac{0.58}{0.42} = 1.38.\end{aligned}$$

This means that content-proficient students with high CRT scores were 1.38 times more likely to answer the target question correctly than content-proficient students with low CRT scores.

As can be seen in the equation above, 58% of high-CRT-scoring content-proficient students answered the target question correctly while only 42% of low-CRT-scoring content-proficient students did so. Therefore, for content-proficient students who demonstrated an understanding of the appropriate content knowledge by answering the screening question correctly, a higher propensity towards cognitive reflection corresponded to a higher probability of success on the target question. In other words, in order to answer the target question correctly, not only did students need particular mindware, but they also benefited from a tendency to overcome an intuitive thought with analytical reasoning. Higher cognitive reflection skills gave content-proficient students an advantage on the target question. It would follow that success on the target question requires some amount of cognitive reflection, and for students with relevant content knowledge who fail to answer the target question correctly, a reasoning error may very well be to blame, consistent with DPTs.

Content-proficient students' CRT scores predict target performance A third method we used to test whether the screening and target questions were identifying students making reasoning errors involved binary logistic regression. The benefit of a binary

logistic regression is that it allows us to see if CRT score is a predictor of target question performance for content-proficient students. We would expect that the higher their CRT score, the greater the likelihood that content-proficient students would answer the target question correctly.

Fig. 2.7 shows jittered data points and the corresponding logistic regression for all versions. Each point represents one content-proficient student's CRT score and performance on the target question (0 = incorrect, 1 = correct). The regression shows the probability of answering the target question correctly as a function of CRT score. It is evident that higher CRT scores correspond to a greater likelihood of success on the target question for content-proficient students. Specifically, for every one-point increase in CRT score, the odds of a content-proficient student correctly answering the target question increased by 1.38 (95% C.I. [1.17, 1.63]).

Fig. 2.8 shows a forest plot of odds ratios with error bars representing the 95% confidence intervals from the binary logistic regressions conducted for each version of the question sequence. Note that while versions 1 through 3 were each given during a single academic quarter, version 4 was given to students during three different quarters; therefore, on the forest plot, results from version 4 are also disaggregated by the quarter in which it was given. Since the sample sizes are fairly small for each quarter the question sequence was administered, the error bars are quite large. However, there is a general trend towards odds ratios greater than 1, indicating that an increase in a content-proficient student's CRT score corresponds to an increase in the odds of answering the target question correctly.

Given the results from conducting binary logistic regressions, it appears that cognitive reflection is a predictor of success on the target question for content-proficient students.

Discussion of CRT results The results of our analyses are generally supportive of an interpretation based on dual-process theories, but not conclusive, consistent with other studies. In some cases, CRT scores have been associated with certain response patterns on screening-

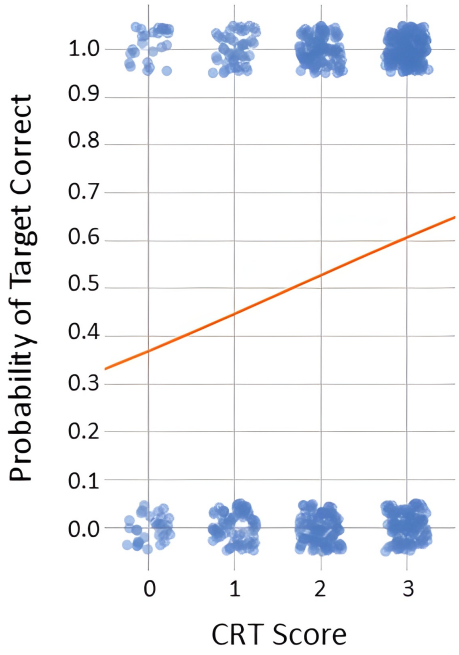


Figure 2.7: Binary logistic regression for all versions of content-proficient students' probability of answering the target question correctly prior to intervention as a function of their CRT score (in orange) ($p = 0.0001$, odds ratio = 1.38, 95% C.I. [1.17, 1.63]) overlaid with jittered CRT score data (in blue). Each blue data point represents a content-proficient student, indicating their CRT score (exactly 0, 1, 2, or 3) along with their performance on the target question (0 = incorrect, 1 = correct). (The data points have been jittered for visualization purposes.)

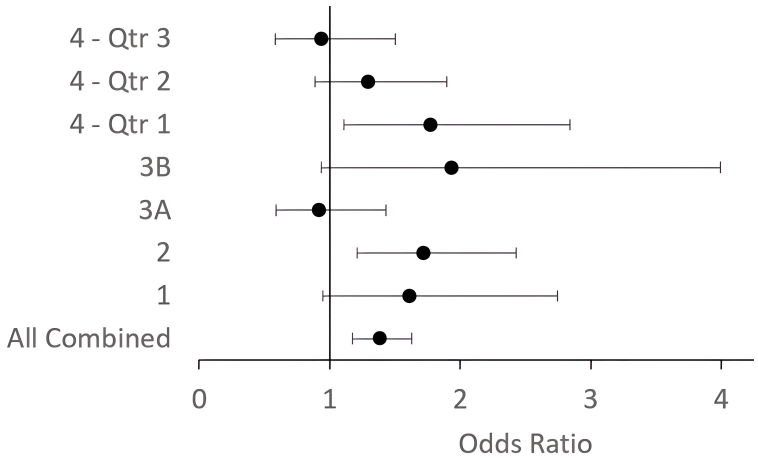


Figure 2.8: Forest plot of odds ratios with 95% confidence intervals for binary logistic regressions on all versions of the question sequence. The vertical axis is labeled by version. Note that version 4 was given in three different academic quarters.

target question pairings while in others no association has been found. For example, Gette and Kryjevskaia [25] conducted a study using one screening question and two target questions related to forces and Newton's third law and found a predictive relationship between CRT score and physics question performance. They observed that students with higher CRT scores were more likely to answer each target question correctly and were more likely to answer all questions consistently. However, Kryjevskaia *et al.* [27] found no association between students' performance on a different target question regarding forces and Newton's laws and their level of cognitive reflection skills. It is possible that some students (even with high cognitive reflection) felt so strongly about their initial incorrect process 1 response that they did not engage in cognitive reflection on the target question.

It is also possible that the CRT is not solely a measure of cognitive reflection, as it correlates with other cognitive and/or demographic measures [21, 36, 37, 38]. Students who have higher CRT scores may generally be successful at answering all kinds of questions. Additionally, the type of cognitive reflection needed on the pulse target question could be different than that needed on the CRT. We might also be witnessing the result of a ceiling effect in CRT scores,

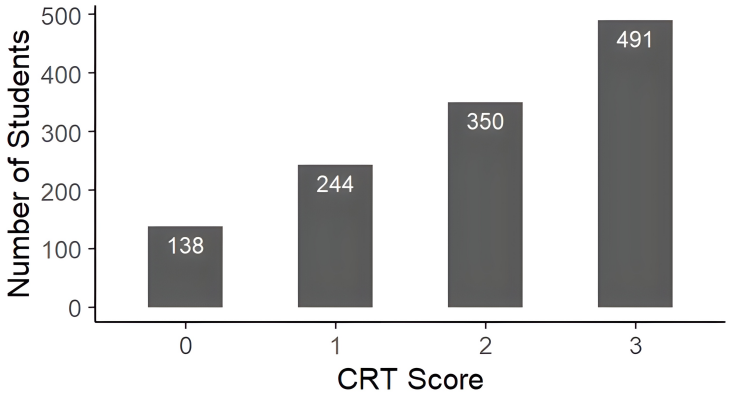


Figure 2.9: Frequency of students earning each CRT score for all versions.

since the most common CRT score for the students in this study was a 3 (see Fig. 2.9).

The small associations we observed between target performance and CRT score for content-proficient students might also be explained by the occurrence of several content-proficient students answering the target question incorrectly even after engaging process 2. In this scenario, these students may not have had strong enough mindware to be able to generate the correct response. Generally, these students may have had a tendency to engage process 2 to further consider an initial process 1 idea, translating to a high CRT score but not translating to a correct answer on the target question in this physics context.

It is also possible that some students found the presence of the CRT questions in a physics course sufficiently incongruous that they were more inclined to reflect than usual, leading them to receive higher CRT scores. These same students might not be so inclined on physics questions, leading to incorrect answers on the target question. In such a case we would expect to see weaker associations between CRT scores and correctness on the target question.

Finally, we note that if students who answered the target question correctly did so because their process 1 result was correct, then their propensity to cognitive reflection would not be relevant. They might then have a broad distribution of CRT scores, making them difficult

to distinguish from students who did not answer the target question correctly.

A brief intervention improves performance on target question

As mentioned earlier, if incorrect answers to the target question by content-proficient students can be attributed to inadequate reflection of answers generated by process 1, and not to inadequate conceptual understanding, then it should be possible for an intervention that promotes reflection to improve performance. In this section we briefly outline an intervention we developed, which is described in greater detail elsewhere [30]. Our intervention was aimed at students who we have reason to believe were equipped with the necessary mindware but failed to effectively engage process 2. Therefore, the intervention was not intended to develop any new content understanding or skills. Instead, it served to steer students away from the intuitive appeal of common unproductive ideas, similar to [26] (but implicitly) and to remind students of physics knowledge they already correctly applied to the screening question.

As outlined earlier, the intervention followed the target and screening questions (see Fig. 2.3). The entire sequence was administered online, and students were expected to be able to complete it in less than 30 minutes.

On the latest iteration, version 4, two fictitious students discuss the target question. They agree that the pulse speed along the right spring was unchanged between experiments 1 and 2 (correct). However, they disagree on the time for the pulse to pass the junction. In this way, students were reminded of relevant mindware while being nudged to doubt the idea that the time to pass the junction was different in experiment 2. The students are then asked how each of these fictitious students would answer the target question. By considering alternative lines of reasoning, students were encouraged to engage in productive analytic thinking (*i.e.*, process 2). A subsequent question asked students to reflect on how (if at all) their answer to the target question would change after considering the fictitious student statements. Students who indicated a desire to make a change had such an opportunity on the next screen of the quiz where the target question was presented a second time.

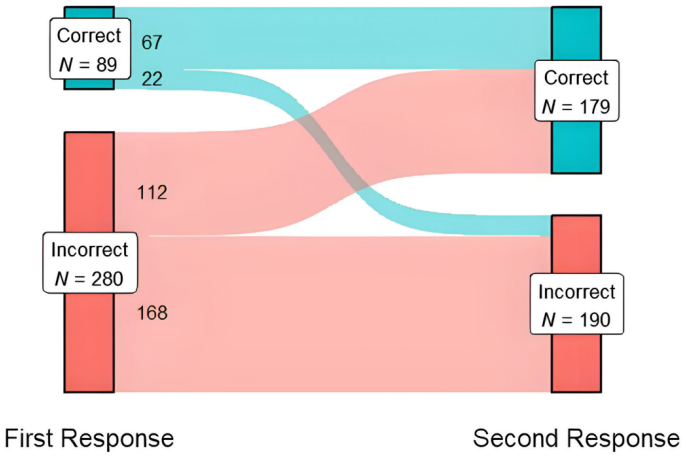


Figure 2.10: Sankey diagram showing pathways of answering patterns on first and second instances of the target question for version 4.

We would expect that after encountering our intervention, student performance on the target question would improve, especially for content-proficient students who made reasoning errors. In fact, for all students who chose to re-answer the target question, we did see better performance. Several students improved accuracy by switching their response from an incorrect choice on the target question to the correct answer when revisiting the target question, while much fewer decreased accuracy by switching from the correct answer to an incorrect one (see Fig. 2.10). A McNemar's test suggests that there is a difference between these two groups of students whose accuracy on the target question changed, with more students improving accuracy [$p = 7.56 \times 10^{-15}$, Cohen's g effect size $= 0.336$ (large)]. Table 2.8 shows a breakdown of student performance for first and second responses to the target question.

Even though the intervention improved success on the target question for all students, not just those we presume made reasoning errors, it is unlikely that student success can be explained entirely by strengthened content knowledge, which was not the intent of our intervention. If students answered the target question incorrectly due to inadequate conceptual understanding, we would expect an intervention that addresses this issue to have the same effect on all students regardless of their performance on the screening question.

Table 2.8: Contingency table of version 4 student performance on the first and second instances of the target question for those who decided to re-answer the target question after seeing the intervention.

	Target second response correct	Target second response incorrect	Total
Target first response correct	67	22	89
Target first response incorrect	112	168	280
Total	179	190	369

Table 2.9: Performance on the second instance of the target question for content-proficient and noncontent-proficient students on version 4.

	Target second response correct	Target second response incorrect	Total
Content proficient	95	46	141
Not content proficient	84	144	228
Total	179	190	369

However, we saw that the intervention was more likely to benefit those students who demonstrated application of mindware by answering the screening question correctly. Student performance when re-answering the target question in version 4 is displayed in Table 2.9, disaggregated by students' content-proficiency. A chi square test indicates that a greater proportion of content-proficient students re-answered the target question correctly compared to noncontent-proficient students [$p = 1.18 \times 10^{-8}$, Cramér's V effect size = 0.30 (medium)].

We also calculated a risk ratio to compare students' likelihood of re-answering the target question correctly based on their content-proficiency using the data from Table 2.9:

$$\text{Risk Ratio} = \frac{\text{Target second response correct \& content proficient/Content proficient}}{\text{Target second response correct \& not content proficient/Not content proficient}} = \frac{95/141}{84/228} = \frac{0.67}{0.37} = 1.81.$$

This ratio indicates that content-proficient students were 1.81 times more likely to re-answer the target question correctly than those who were not content proficient.

Recall that our intervention was designed to prompt process 2 engagement by encouraging students to consider alternative lines of reasoning. Given that, after the intervention, student performance on the target question improved, especially for content-proficient students, this data supports the hypothesis that, for content-proficient students, reasoning that produces correct answers to the target question can be explained by effective engagement of process 2. These results are consistent with prior research using different intervention strategies [12, 28].

2.7 Conclusion

Dual-process theories suggest that there are two processes involved in reasoning and decision-making: the automatic, intuitive process 1 and the deliberate, rule-based process 2. An unproductive idea generated by process 1 can be overridden by the effective engagement of process 2. Under this framework, it is possible for a student to be capable of applying relevant content knowledge on a particular physics question yet rely on a misleading initial response generated by process 1, resulting in an incorrect response. Previous research proposed these theories as an explanation for observed inconsistencies in student responses to questions that ostensibly test the same concepts.

Though there are nuances, the data presented here generally support this perspective. We collected data from a screening and target pair, a set of related questions of our own design, a brief intervention that was based on dual-process theories, and the cognitive reflection test.

Our results generally support assumption 1: the screening question is generally identifying

students who have the mindware needed to answer the target question. As discussed above, this assumption implies that (i) students who answer the screening question incorrectly are not able to answer the target question correctly, and (ii) students who answer the screening question correctly should also be able to answer related questions that also lack salient distracting features.

For point (i), our results confirmed that students who answered the target question correctly were more likely to have also answered the screening question correctly. For point (ii), we found that content-proficient students were more likely than noncontent-proficient students to correctly answer the "bug" questions, which tested an aspect of conceptual understanding needed to answer the target question more directly. Of course, no single test can clearly distinguish between those who have mindware and those who don't (a binary that is itself an oversimplification) and there will be some number of students who were misidentified by the screening question and whose incorrect responses to the target question stemmed from inadequate conceptual understanding.

Our results also broadly support assumption (2): among content-proficient students (those who answer the screening question correctly), responding to the target question is largely an issue of cognitive reflection on misleading process 1 answers. This assumption implies that (iii) content-proficient students who exhibit a higher propensity for cognitive reflection should be more likely to answer the target question correctly; and (iv) interventions designed to support students in reflecting on their thinking, without promoting further conceptual development, should increase the number of correct responses to the target question.

For point (iii), we examined whether content-proficient students who answered the target question correctly had a generally higher propensity to engage in cognitive reflection as measured by the CRT. On all versions of our question sequence, there were only small differences in the frequency of high and low CRT scores between all-correct students and reasoning-error students. However, content-proficient students with high CRT scores were more likely to answer the target question correctly than their low-CRT-scoring counterparts,

and CRT score was generally a predictor of target question performance for content-proficient students.

The nature of cognitive reflection test scores in relation to content-specific screening-target question responses needs further investigation as there are several open questions: Is cognitive reflection as measured by the CRT the same kind of cognitive reflection needed on various target questions? Is the CRT an independent measure of cognitive reflection or are students with high CRT scores also higher-performing students in general who are likely to answer any questions well? Since the most common score on the CRT for the population of students in this study was a 3, is there a ceiling effect that is causing the CRT to ineffectively discriminate between students' cognitive reflection skills? We have started investigating these issues to gain a better understanding of the CRT and its relation to our research.

For point (iv), our study showed that the intervention questions we developed based on DPTs had a positive impact on student performance on the target question, particularly for content-proficient students. We would not expect to see this kind of impact if student inconsistencies could merely be explained by conceptual difficulties.

2.7.1 Limitations

It is important to note that our data is limited to a single population consisting of introductory physics students in calculus-based physics at a single institution, which in many respects may not be representative of introductory physics students more broadly. One potential challenge to generalizability, at least within the United States, is the demographic distribution of the students [39]. We offer the data below in acknowledgment of this possibility, and to support further analysis. A breakdown of student-reported ethnicity is shown in Fig. 2.11 for the participants for whom we had access to such information ($N = 1{,}179$). Although the categories are not the same, and our sample is limited to STEM students in calculus-based physics, comparing this data to national data [40] indicates that our sample has a greater proportion of Asian and International students, and smaller proportions of Black, White and

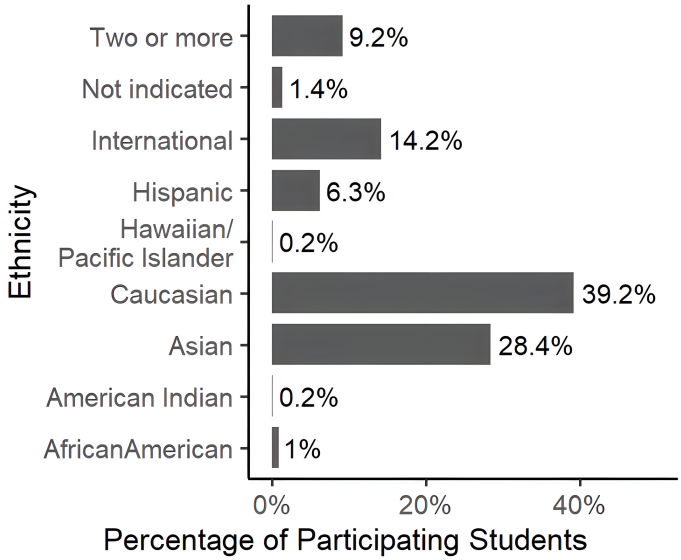


Figure 2.11: Self-reported ethnicities for the majority of student participants.

Hispanic students than the national population of college freshman.

It is also important to note that the focus of our analysis was on content-proficient students, the number of whom ranged from 44 to 140 during each quarter and individual version, so some of our samples were relatively small.

2.7.2 Further research

Our data indicates that the pulse question sequence seems to be accurately distinguishing students' conceptual understanding and reasoning abilities. It appears that the screening question is more-or-less assessing student mindware, and the screening-target question pair is identifying students making reasoning errors to some extent. However, there are several open questions regarding nuances in the methods used to test the effectiveness of the screening-target question functionality. This is a topic that needs further investigation, and our study provides a window into areas where specific research is needed.

The use of screening-target question pairs has been used in multiple studies to disentangle

students' conceptual understanding and reasoning strategies under the framework of dual-process theories [24, 29, 27, 25, 26, 12, 28], but more work needs to be done to validate such question pairings. Further research needs to be completed that evaluates the cognitive reflection test and its relation to screening-target pairings. In the current study, the 3-item CRT was used; however, a 7-item CRT also exists which may do better at discriminating between students' cognitive reflection skills. This more extensive CRT will be used and analyzed in relation to question sequences involving physics content that includes the pulse-on-a-spring questions in addition to other topics. Previous research suggests that there is a correlation between CRT scores and other academic measures such as SAT scores [21, 37, 38]. We are in the process of examining the degree to which the CRT can still be useful for identifying students with a strong tendency to reflect, and not simply general academic prowess. Another avenue of research under investigation is to see what associations there may be between CRT scores and responses to a variety of screening-target pairs with the intent to uncover information about why there are correlations for some pairs and not others.

2.7.3 Implications for instruction

Instructors should be aware that while some students may have adequate content knowledge, they also need cognitive reflection skills to be able to succeed on a variety of physics questions. Understanding the human reasoning process as described by dual-process theories can make this clearer for instructors and students alike. Explicit instruction on DPTs could potentially help students become more aware of their own thinking processes and reduce possible feelings of inadequacy due to misleading intuitions about physics. Additionally, encouraging students to test their intuitions using process 2 as part of the classroom culture could help students overcome reasoning errors.

2.8 Acknowledgments

The authors would like to recognize the many contributions of Andrew Boudreaux, Mila Kryjevskaia, Beth Lindsey, Drew Rosen and MacKenzie Stetzer to the study described in

this paper. Special thanks are also due to Andrew Heckler from the Ohio State University for his input on data analysis. We would also like to thank the members of the University of Washington Physics Education Group for their feedback during the process of writing this paper and two anonymous reviewers for their helpful comments. This material is based upon work supported by the National Science Foundation under Grants No. DUE-1821390, DUE-1821123, DUE-1821400, DUE-1821511, and DUE-1821561.

Chapter 3

AN INTERVENTION TO HELP STUDENTS RECOGNIZE AND RESOLVE REASONING INCONSISTENCIES: AN APPLICATION OF DUAL-PROCESS THEORIES

Kristin Kellar and Paula Heron

Department of Physics, University of Washington, Seattle, Washington 98195

3.1 Abstract

Dual-process theories of reasoning suggest that humans reason using two processes, sometimes referred to as "intuitive" (process 1) and "analytic" (process 2). When students are faced with a physics question, process 1 may produce a quick and appealing solution. If the student reflects on that answer and engages process 2, and if they have adequate content knowledge for a valid analysis, they may find their initial solution to be faulty. If they don't reflect, or if their content knowledge is inadequate, they may not detect the fault in their initial solution. Physics education researchers have been investigating this model, in part by attempting to tell whether incorrect answers stem from inadequate reflection or inadequate content knowledge. Researchers are also attempting to help learners reflect on, and possibly correct, their initial solutions. This study shows how a short online intervention can help introductory physics students recognize reasoning inconsistencies and activate process 2 effectively to override initial incorrect responses on a particular task.

3.2 Introduction

For several decades, researchers have investigated how introductory physics students respond to questions that require qualitative reasoning. In particular, researchers have sought to explain why students who have been taught the relevant concepts often answer incorrectly. In many cases, errors can be traced to inadequate conceptual understanding. Researchers and teachers have also observed that even students who demonstrate an adequate grasp of the relevant concepts by applying them correctly on some questions can answer other questions incorrectly, a phenomenon physics education researchers have recently explained in terms of dual-process theories (DPTs) [24, 29, 27, 22].

DPTs vary in several respects, but generally suggest that human reasoning relies on two processes: one sometimes referred to as "intuitive" (process 1) and one sometimes referred to as "analytical" (process 2) [19]. Kryjevskaia *et al.* [24] used DPTs to explain a pattern of responses to a pair of questions about pulse propagation: a "screening" question intended to identify students with adequate knowledge of the relevant concepts, and a "target" question that tests those concepts in a situation with distracting features. We subsequently examined responses to these questions, others of our own design, and a three-item test used to assess one's tendency towards cognitive reflection known as the cognitive reflection test (CRT) [41, 21]. We found support for the interpretation offered by Kryjevskaia *et al.* [24]. Part of the evidence we presented came from the results of a self-contained, written intervention we designed to help students reflect on their responses to the pair of questions, and engage in analytical thinking. In this paper we describe the development of that intervention and our findings in greater detail. Thus this study adds to the literature that corroborates the use of DPTs to explain observations in physics classrooms and to guide the development of instructional strategies.

An important recent contribution to that literature is a paper by Lindsey *et al.* [28]. The authors examined responses to tasks that require students to apply Newtons' second law to

compare the drag forces on two objects at their terminal velocities. The authors developed two interventions, using different strategies to spur students to reflect on, and correct, their initial answers. Their study, and the one we report here took place in parallel, with frequent conversations among all authors. However, we employed somewhat different tactics, addressed a different topic area, and operated in a different instructional context. Moreover, we employed a different assessment strategy. Our study can thus be viewed as complementary to theirs. Both studies form part of a larger project that aims to establish a framework for curriculum development informed by DPTs, in part by linking specific tactics to specific reasoning challenges.

This paper starts by briefly outlining the underpinnings of dual-process theories, and relevant prior research. We then describe the iterative process for intervention development and present results. These results are discussed in the context of theory and other research, and implications for instruction and further research are provided.

3.3 Theoretical Framework

A common feature of DPTs is their proposal that humans employ two processes, or systems, to reach conclusions or make decisions. One process, variously called "fast," "heuristic," or "intuitive," is believed to operate in the background, providing a first impression in response to a prompt of some kind. If something about that first impression seems to warrant closer scrutiny, another process, variously called "slow" or "analytic" comes into play. That process is effortful, and according to DPTs, avoided if possible - a phenomenon referred to as *cognitive miserliness* [20] or *cognitive frugality* [22]. When either process is followed, but especially process 2, adequate *mindware*, or knowledge of the rules, concepts and relevant principles, is needed [20]. In the case of many of the tasks that initially motivated the development of DPTs, that mindware is assumed on the part of most adults. Errors made are thus attributable to reasoning faults. In the case of physics tasks presented in a physics course, that mindware is in the process of being developed. Therefore distinguishing between

reasoning faults and inadequate mindware is challenging.

DPTs have been used to explain patterns of responses, and have guided the development of interventions designed to help students answer conceptual questions that are made more challenging by the presence of *salient distracting features* [14, 32, 33]. These features can influence the formation of the first impression, that if especially compelling, can preclude any further analysis. The framework summarized in Fig. 3.1, which we employ in our work, represents this process as the *path of cognitive frugality*. Specifically, the formation of an initial response that is intuitively appealing but incorrect or misleading (hazard A) followed by inadequate reflection (hazard B) leads swiftly to an incorrect conclusion. The incorrect response may obscure the presence of conceptual understanding because that understanding was simply never called upon.

Reasoning hazard B can be referred to as *detection failure* [20] since one does not recognize a conflict between an intuitive response and learned normative rules. If one avoids hazard B by detecting the need to reflect on the model generated by process 1, process 2 is engaged on the reasoning path known as the *path of sustained effort*. If fruitful, a new correct model is generated that overrides the initially incorrect model. However, while on the path of sustained effort, it is possible to use process 2 to validate the provisional model through confirmation bias instead [23] (hazard C), and so an incorrect response is maintained.

The framework represented in Fig. 3.1 assumes that *cognitive reflection* is responsible for shifting from process 1 to process 2. The propensity toward reflection is often assessed by the popular three-item cognitive reflection test (CRT), shown in Fig. 3.2. The questions on the CRT require minimal mindware and tend to elicit an incorrect first response. Therefore, one must activate process 2 and take the path of sustained effort to override an initial incorrect answer with a correct conclusion. Scores range from 0 to 3 depending on how many questions one answers correctly. A score of 2 or 3 is considered "high," and indicates a higher propensity towards cognitive reflection while a score of 0 or 1 is considered "low," and corresponds to more miserly thinking.

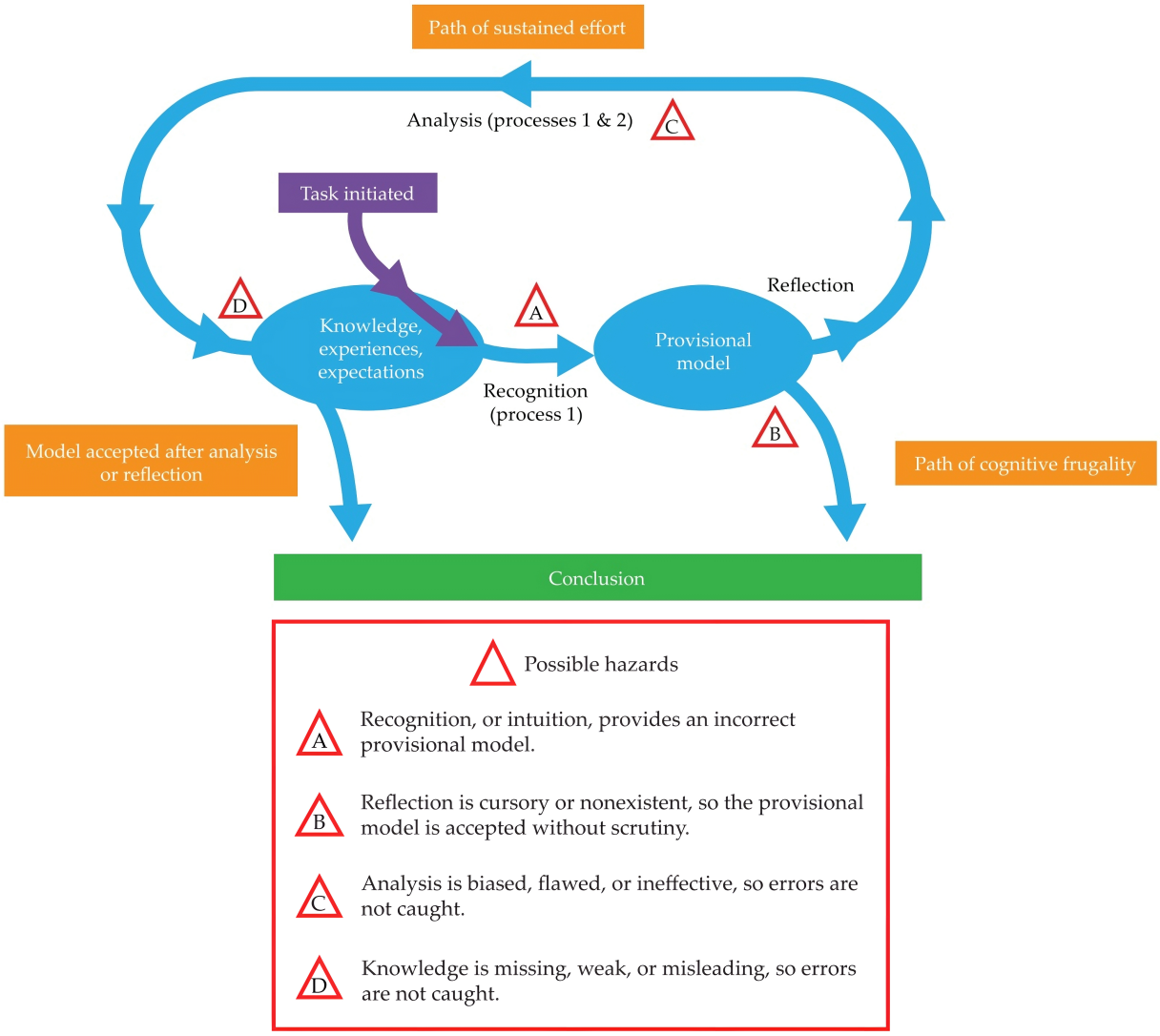


Figure 3.1: Dual-process theories of reasoning suggest two main thinking processes. When presented with a task, these processes can activate and interact via various reasoning pathways to arrive at a solution. Diagram from [22].

(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? _____ cents

(2) If it takes 5 machines to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes

(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 of the lake? _____ days

Figure 3.2: The cognitive reflection test, developed by Frederick [21].

3.4 Prior Research

A number of studies have used DPTs in the context of physics teaching. We discuss the most relevant research here and how it relates to the study described in this article.

A study by Kryjevskaia and colleagues [24] motivated the current work. The authors described a pair of questions about pulses. (We describe it briefly here and in greater detail below.) The students were presented with a scenario involving two connected springs and told that in "experiment 1," a student creates a pulse on the left spring, which is transmitted to the right spring. Then they were told that in "experiment 2," a change has been made that doubles the width of the pulse on the left spring. They are told that neither the motion of the hand nor the tension in the springs changes and that the right spring remains unchanged. The "screening" question asks students to determine the change that was made in experiment 2; the "target" question asks them to compare the width of the pulse on the right spring in the two experiments. Many students who correctly conclude that the medium of the left spring must have changed incorrectly conclude that the pulse on the right spring must be wider in experiment 1 "because the incident pulse is wider." The authors hypothesized that students who answered the screening question correctly had the

conceptual understanding necessary for answering the target question, so if they answered it incorrectly, then they were likely relying on a misleading process 1 response.

In a previous study [41] we examined this hypothesis in detail. We found that students who answered the target question correctly were likely to have answered the screening question correctly, that students who answered the screening question correctly were likely to answer other straightforward conceptual questions correctly, that students with higher scores on the CRT were more likely to answer the target question correctly, and that a short intervention (reported on here) that did not intend to bolster conceptual understanding increased performance on the target question. We concluded that this evidence generally supports the hypotheses of Kryjevskaia *et al.* [24].

Another paper by Kryjevskaia *et al.* [27] reports on an investigation of student responses to a set of tasks that involve static friction, and an intervention intended to help them apply Newton's second law, rather than focusing solely on rules about frictional forces of limited applicability. In a target question, students were asked to determine the magnitude of the frictional force exerted on a magnet on a fridge when an upward force is exerted on the magnet, which remains at rest. The given magnitudes of the weight and applied force can be used to infer that the frictional force is directed upward, in the same direction as the applied force. Many students responded instead that the frictional force must be directed downward because "friction opposes applied forces." The intervention involved three stages consisting of both individual and group work on various questions related to the tasks and Newton's second law. The intervention improved performance on a post-test question and was more effective for students with higher CRT scores.

In Lindsey *et al.* [28], the context involves two disks that have the same mass but different radii. Students are told that the two disks are "released from rest and allowed to fall through the air. Each object speeds up until it reaches a constant speed, its terminal speed." In the "target" question, students are asked to compare the drag forces on the two disks once they each have reached terminal speed. A correct answer can be obtained by recognizing that at

constant speed, the net force on each disk is zero, and since the downward gravitational forces on each disk are equal, the upward drag forces must also be equal. Many students answer incorrectly that the drag force on the object with a larger cross-sectional area is greater. The authors then tested two interventions in a controlled experiment. Both interventions began with a series of questions intended to lead students to conclude that the net force on each object is zero, and to recognize that identical free-body diagrams can be drawn for each one. The "cognitive" intervention attempted to support the appropriate reasoning that leads to a correct answer. In short, this style ignores the intuitive answer and the features that suggest it, in favor of reinforcing the correct reasoning exhibited on the initial questions. In contrast, the "metacognitive" intervention takes direct aim at the intuitive response, leading students to recognize its incompatibility with the reasoning needed for the screening task. Most importantly, this intervention leads students to reconcile the (correct) principles that support an incorrect model with the (correct) principles that support a correct one. This process affirms the relevance of the formula for drag force, and shows how it can be used in concert with Newton's second law to draw inferences. Thus the students' intuition about the cross-sectional area being relevant is not refuted, rather it is brought into alignment with the conceptual framework they are being taught. The metacognitive approach was found to be more effective at increasing performance on the target question, and on a separate aligned question within the same assignment.

The friction [27] and air-resistance [28] cases have some common elements. The heuristics "friction opposes an applied force" and "greater area implies greater drag force" are reminiscent of specific rules presented in class, and, at least in the case of the drag force, were expressed mathematically. The "physics-like" nature of these heuristics may give intuitive process 1 answers a veneer of legitimacy and inhibit serious scrutiny. Moreover, in both cases, correct analysis depends on the notoriously difficult Newton's second law. More generally, a correct analysis requires recognizing a hierarchy of concepts in which formulas such as $F_{drag} = 1/2\ \rho v^2 C_{drag} A$ rank lower than Newton's laws. The strategy of *reconciliation*

involves affirming the relevance of the equations while leading students to recognize and question an implicit, unfounded assumption. In the case of the air resistance target question, the assumption students make is that the disks are moving at the same speed. (The problem statement does not specify whether they are or not, just that both have reached terminal speed.) Even students who correctly compare the disks' terminal speeds on an earlier question abandon that comparison when asked about the drag forces. After students are led to recognize that Newton's laws require that the drag forces must be the same, they are led to recognize that the drag equation can be used in tandem with that fact to reach a conclusion about the relative terminal speeds. In this way the intuitive recognition that surface area matters is reconciled with the analytical conclusion that the drag forces are the same. The tactics used include a series of scaffolding questions and a direct challenge to the intuitive response in which students are shown a statement that likely resembles their own thinking, to which they are asked to respond.

In the case of the pulse questions that are the subject of this article, the dominant heuristics used by students relate the width of an outgoing pulse to the width of the incoming pulse directly, rather than indirectly through their mutual dependence on the time taken to generate the original disturbance. Students have not been taught a mathematical expression that includes pulse width, and the equation v=f is not obviously relevant to the problem at hand. In the classes in which these students were enrolled, discussions of junctions between different media tended to focus on continuity of frequency and, where relevant, tension. They may recognize, correctly, that for a given pair of media, an incident pulse of greater width (or an incident wave of greater wavelength), gives rise to a transmitted pulse of greater width (or a transmitted wave of greater wavelength). Given that a pulse's width and height are likely its most visually salient features, the support for the heuristic may be sufficiently compelling that no further scrutiny occurs. The actions that can be taken to manipulate these features, in effect the causes for the observed effect, are less salient, and thus may play a secondary role, if any, in student thinking. In fact, the heuristic may reflect even more basic ideas:

upstream changes have downstream consequences or even simply large begets large.

The reasoning required to analyze the situation properly requires not so much a hierarchy in which special cases rank lower, but a recognition of what is a cause (initial disturbance) and what is an effect (pulse width, or wavelength). In other words, the structure of the heuristics at play in the pulses context differ from those seen in [27] and [28]. In fact, the heuristics differ in both generating provisional models via process 1 and the reasoning required to obtain a correct response via process 2 (unless a level of expertise is reached at which process 1 itself produces a correct initial model). Therefore we argue that the tactics needed in an effective intervention in the pulses context may differ from those featured in [27] and [28], even when adopting the broad principle of directly addressing common incorrect process 1 answers.

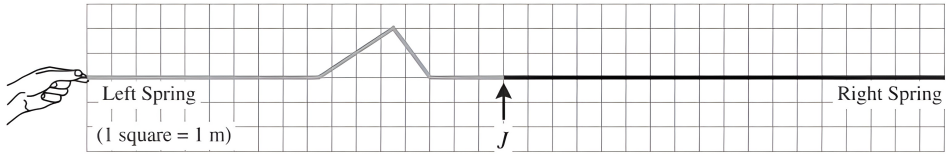
3.5 Overview of Study

The intervention we present was given to students as part of a question sequence and was developed iteratively. We describe several versions here in an effort to pinpoint which aspects contribute to its effectiveness. Each question sequence included the same screening-target pair. Each sequence also included questions intended to encourage reflection and the engagement of analytic thinking processes. In all cases, students were offered an opportunity to revisit, and then reanswer the earlier target question. Below we describe the basic components of the sequences and the reasoning behind their design.

3.5.1 Focus of intervention

The focus of this study is the pair of screening and target questions about pulse transmission described briefly above. We adapted these questions for easier analysis by dividing the target question into two separate parts, one multiple choice and one short answer, as shown in Fig. 3.3. After answering, students are instructed to make a note of their answers to these questions before continuing on to the next page of an online quiz where they could not go back.

In experiment 1, two different springs are connected at a junction point *J*. A student generates a pulse on the left spring shown below. It takes the student's hand a time Δt_0 to quickly move the end of the spring back and forth in order to generate the pulse. In experiment 1, the propagation speed of a pulse on the left spring is 1.5 times that on the right spring (v_L=1.5v_R).



Experiment 2 is nearly identical to experiment 1 except for a ***single change.*** As a result of this change, the width of the generated pulse (in the left spring) is doubled. The tension in the spring on the left and the time it takes for the student's hand to move to create the pulse is the same in both experiments. The spring on the right is unchanged.

1. Determine the change that has been made in experiment 2. Explain.

2. Is the width of the ***transmitted pulse*** on the right spring in ***experiment 2*** *greater than, less than,* or *equal to* the width of the ***transmitted pulse*** on the right spring in ***experiment 1***?

3. Explain your response to the previous question.

Figure 3.3: Screening and target questions adapted from [24].

In order to answer the screening question correctly, students need to understand that pulse width depends on (1) the time taken to create the pulse and (2) the speed of pulse propagation, which in turn depends on the linear mass density of the spring and its tension. Because the time taken to create the pulse has not changed, nor has the tension, it can be inferred that the left medium itself has been changed. Specifically, given that the pulse width has doubled, it can be inferred that the propagation speed has doubled, and therefore that the linear mass density in the left spring has decreased by a factor of four. If students responded by stating that the medium of the left spring must be changed (regardless of whether or not they specified how the medium changed), their answers were considered correct since this type of response exhibited the minimum mindware needed to answer the target question correctly. The correct response to the target question requires the same mindware. Students can recognize that given there has been no change in the tension or the medium of the right spring between experiments, the speed of the pulse would not change. As the time to generate the pulse remained constant, the width of the pulse would be equal in both experiments. In our previous paper we presented evidence that students who answer the screening question correctly tend to have adequate conceptual knowledge for answering the target question and that many incorrect responses to the latter can be attributed to failure to reflect.

3.5.2 Intervention strategy

Research suggests that students will not abandon an unproductive intuitive model unless they have cause for doubt in their initial model and they have the content knowledge necessary to produce the correct model [20, 42, 43]. Therefore, an effective intervention question sequence should (1) raise enough doubt in an unproductive process 1 response to the target question that students reflect and engage process 2, (2) remind them of the relevant knowledge and skills needed to successfully navigate the path of sustained effort, and (3) encourage analytic thinking so that students generate a new provisional model, ultimately leading to a correct response to the target question.

For some students, it is plausible that interventions that raise doubt in an initial intuitive answer, or simply inject a pause into the process, might be sufficient. Once engaged in analytical thinking, equipped with the content knowledge that allowed them to succeed in the screening question, a successful override of their process 1 answer could result. However, it appears that for many students, support for the analytical process needs to be more direct and sustained. Therefore our strategy has two main goals: to engage students in process 2 reasoning and to support them in doing so. In the terms of the framework described in Fig. 3.1, this means diverting students onto the path of sustained effort, and helping them avoid hazard C, in part by foregrounding the knowledge needed to complete the process successfully.

Devising the specific features of an intervention involves careful analysis of the reasoning used by students and the reasoning needed to correctly answer the target question, with an eye to identifying areas of overlap. In the pulses case, a correct response to the target question requires recognizing that the medium determines how fast the pulse reaches the junction whereas the hand motion determines how fast it passes the junction - changing either will affect the pulse width as long as tension remains the same, as was indicated in the question. Given that the question statement specifies that neither the medium on the right, nor the motion that created the pulse changes, the answer that the width of the pulse on the right doesn't change should be easily found. However, for the screening question, it is possible, in principle, for students to correctly answer without recognizing the connection between the hand motion and pulse width. Specifically, they don't necessarily need to recognize that the hand motion, if changed, could be responsible for the observed increased pulse width on the left spring. We don't have any evidence that many students were assuming, even implicitly, that the motion of the junction had changed when the incident pulse width changed, but we believed it was likely important to elicit thinking about the junction motion in order to reach a valid conclusion about the transmitted pulse width. Thus part of our strategy would be to raise the accessibility of this knowledge.

Our intervention was intended to support students who made reasoning errors on the target question by helping them to reflect on that question and effectively engage in process 2 thinking. Therefore, our strategy has three components based on the research described above: (1) raise doubt in an initial incorrect model, (2) cue relevant knowledge needed on the target question (and implicitly needed for the screening question), and (3) encourage productive analytic thinking.

The first component is necessary, though not sufficient. If students have no reason to doubt their initial model, they may not engage in the process 2 thinking at all. If the intervention was able to produce such awareness of the necessity for taking the path of sustained effort on the target question, it would allow students to overcome the hurdle of detection failure (reasoning hazard B) when given an opportunity to revisit the target question.

The second and third components of our strategy were designed to support students once they have entered the sustained-effort reasoning pathway. On this path they need to be able to develop a new mental model, which requires applying relevant mindware and productive analytic thinking. Thus the intervention needs to cue relevant mindware and encourage students to engage in alternative lines of reasoning that test the provisional model (avoiding hazard C) in order to develop a new provisional model.

3.5.3 Structure of interventions

The overall structure of each intervention, administered online to individual students, was the same: screening and target questions followed by (or interspersed with) additional questions intended to stimulate reflection and analysis, followed by an opportunity to detect the need to revisit the target question and an opportunity to reanswer the target question if relevant, as illustrated in Fig. 3.4. Depending on the version of the question sequence used, the "revisit" question either asked students about their consistency between their responses to the target question and a set of additional questions or asked if they wanted to reflect on their thinking about the target question after answering a set of additional questions. In

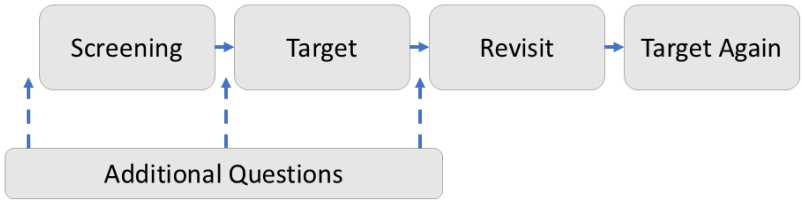


Figure 3.4: The main components of our question sequence.

each case, only those students who recognized inconsistency or indicated that they would like to change their answer and/or explanation to the target question were advanced to a new page in the online environment where they could reanswer the target question.

3.5.4 Instructional context and data collection

This study was conducted in several large enrollment introductory calculus-based physics courses at the University of Washington (UW) during six quarters starting in Autumn of 2020 and ending in Spring of 2022. Student participants were included in the study if they did not indicate that they wanted to opt out. The courses were the third part of a three-course sequence and included material on waves and optics. The course structure contained lecture, lab, and tutorial components. The question sequences discussed here were presented at the beginning of online quizzes called "tutorial pretests" given after lecture instruction, but before small-enrollment tutorial sections on the relevant topic. In general, these quizzes were typically given each week and were available online for two and a half days over the weekend. The time to complete each quiz was limited to 15 min, which was more than enough time for most students. On the quiz in which our question sequence was used each quarter, the time limit was extended to 30 min. Students earned a completion grade for each quiz whether or not they answered the questions correctly, and this amounted to a very small percentage of their total grade in the course. Given that the online quizzes were a regular part of the course structure, including the question sequence on one of the quizzes during a quarter allowed for minimal disruption to the class. Due to the COVID-19 pandemic, the

courses were taught virtually during four of the quarters.

We also administered the cognitive reflection test, shown in Fig. 3.2 [21] on separate online quizzes. If students completed the CRT more than once, since it was administered every quarter of the introductory calculus-based course sequence, their first score was used for analysis. In our previous paper we showed that for students who correctly answered the screening question, CRT scores were associated with performance on the target question. The results are consistent with the hypothesis that some, or many, incorrect answers on the target question are due to lack of adequate reflection. The interventions we designed were specifically intended to aid such students.

3.5.5 Assessment strategy

We developed a series of versions, assessing each one according to the criteria below, and modifying them accordingly. The main goal was to determine if each version was effective, and if so, more effective than the previous version. Although we did not conduct randomized trials, we can use student data to assess whether comparisons between results obtained in a given course or academic term are valid. Prior research regarding introductory calculus-based tutorial pretest data at the University of Washington has shown that results across academic quarters are "essentially the same" regardless of differences in class composition and instruction [34], and so we consider results from each version to be more-or-less comparable across academic quarters.

First, if the intervention as a whole helps students arrive at a correct conclusion, then we expect a greater fraction of students to switch their answers to the question from incorrect to correct than vice versa. If, on the other hand, the intervention operates primarily as a hint to students that they may want to change their answers, we should see roughly the same movement in both directions.

Second, if the intervention operates as intended, *i.e.* it primarily serves to encourage students with adequate content knowledge to slow down and consider the situation carefully, then we

expect to see a greater effect among content-proficient students. In our previous paper we provided evidence that suggests that success on the screening question is a valid measure of content proficiency. Therefore we expect to see students who answer the screening question correctly benefit more than students who do not. If on the other hand the intervention operates in some other fashion, *e.g.*, by bolstering content knowledge, then this effect may not appear.

Third, if the success of the intervention hinges on providing students with support in using their (perhaps tacit) understanding of the significance of the hand motion in determining pulse width, then we expect to see an advantage for students who correctly answer additional questions aimed at promoting analytic thinking.

Fourth, in our previous paper we demonstrated that among content-proficient students, those with higher CRT scores were more likely to answer the target question correctly than those with lower CRT scores. We interpret this to mean that some fraction of students who answered the target question correctly did so by reflecting on an incorrect process 1 answer (avoiding hazard B) and engaging process 2 successfully. Therefore, it is possible that by inducing reflection, the intervention would benefit some students with low CRT scores. It is also likely that some students who reflected spontaneously were unable to sustain an override of their process 1 answer and did not end up answering correctly (stymied by hazard C). Therefore it is also possible that students may benefit from being supported navigating the path of sustained effort regardless of their tendency to reflect spontaneously, in which case no strong association between eventual success and CRT scores should be observed.

3.6 Question Sequence Iterations

In this section we present an overview of the iterative development process, summarizing the first two versions briefly. Then we describe the third version in detail and illustrate how it motivated changes that resulted in the fourth and final version. We present the results of the fourth version in detail.

3.6.1 Summary of versions 1 and 2

Versions 1 and 2 contained different attempts to alert students to potential inconsistencies between their answers to the screening and target questions. In each case, a set of additional questions about bugs resting on the hand and spring from the screening-target setup was included after the screening-target pair. Each iteration of the "bug questions" was intended to draw student attention to the fact that the motion of the junction mimics the motion of the hand. (Note that on versions 1, 2, and 3, the bug questions were very similar, only modified for clarity and conciseness. Fig. 3.5 shows the revised questions used on version 3.) The idea was that even students who answered the screening question correctly may not have fully considered the significance of the hand motion. We asked students if their answers to the target question and the bug questions were mutually consistent, and then offered them the chance to revise their answer to the former if they indicated inconsistency in their answers. These students were directed to a new page of the online quiz where they could revisit the target question including multiple choice and explanation components, after which point they moved on to the rest of the regularly scheduled quiz, which was part of the course and not this study. In all other cases, students were directed straight to the rest of the online quiz as designed for the course. The results showed that these interventions were not successful in significantly inducing students to change their answers on the target question.

3.6.2 Version 3 design

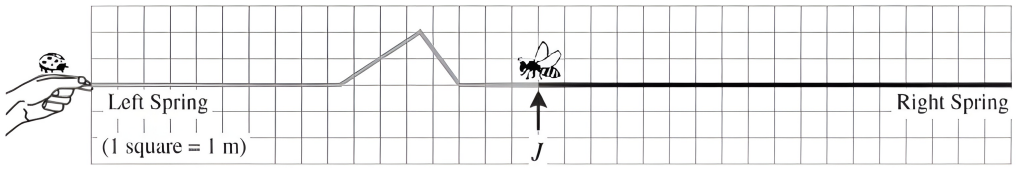
In versions 3A and 3B, the general content describing the physical scenario was updated to include descriptions of the ladybug and bumblebee. The results from versions 1 and 2 suggested that the presence of the bug questions alone would be unlikely to change performance on the target question, but it was possible that seeing these questions prior to considering the target question, rather than after would be more impactful. In short, it seemed possible that the bug questions could cue knowledge useful for the target question, but once that question had been answered, they were insufficient to induce students to change their answers.

Therefore the bug questions were inserted immediately before the target question (3A) or immediately after (3B) so that we could test this possibility. There were two different course sections in Spring of 2021 when this was administered, and each section received one of the two versions.

Asking about consistency between students' responses to the target and bug questions did not seem to induce sufficiently deep reflection and analysis in versions 1 and 2. Therefore version 3 also included pointed *analytic support* questions presented after screening, target, and bug questions in which three fictional students discuss the target question, as shown in Fig. 3.6. As with the drag force example [28], students were presented with incorrect statements that likely mirror their own thinking in a direct attempt to help them identify any flaws.

The first student expresses an argument consistent with the phenomenological primitive "more means more" [44], simply asserting that because the pulse on the left spring has doubled then the pulse on the right would as well. The second student echoes the conclusion, but offers a mechanistic explanation: a wider pulse would take longer to pass the junction. This argument acknowledges that the motion of the junction influences the outgoing pulse width but neglects another variable, pulse speed. (If the speed of the pulse on the left spring had remained the same, a wider pulse would take longer to pass the junction). Fictional student 3 counters by linking the time for the pulse to pass the junction to the (unchanged) hand motion and notes that the motion of the bee would be the same in both experiments. Students were asked if they agreed with any of the fictional students, and to explain. They were then asked whether their response to the question regarding the student dialogue was consistent with their response to the target question. Note that in the previous two versions, students were asked if their responses to the bug questions were consistent with their response to the target question.

Recall the original experiment 1 from the previous page in which two different springs are connected at a junction point *J*.



Suppose that in the original experiment (before any changes had been made) that a ladybug had been resting on the student's hand and a bumblebee had been resting at the junction *J*. Both insects are lightweight and have no effect on the pulse.

Recall that in the original experiment, it takes the student's hand a time Δt_0 to quickly move the end of the spring up and down in order to generate the pulse, and the propagation speed of a pulse on the left spring is 1.5 times that on the right spring ($v_L=1.5v_R$).

1. In ***experiment 1***, will the time it takes the bumblebee to move up and down be greater than, less than, or equal to the time for the ladybug to move up and down in this experiment?
2. Explain your response to the previous question.

Just like on the previous page, experiment 2 is nearly identical to experiment 1 except for the ***single change***. Recall that as a result of this change, the width of the generated pulse is doubled. The tension in the spring on the left and the time it takes for the student's hand to move to create the pulse is the same in both experiments. The spring on the right is unchanged.

3. In ***experiment 2*** after the change is made, will the time it takes the bumblebee to move up and down be greater than, less than, or equal to the time for the ladybug to move up and down in this experiment?
4. Explain your response to the previous question.

Figure 3.5: Additional questions referred to as the "bug questions" that appeared on the version 3 question sequence [41]. Versions 1 and 2 included similar bug questions, which were then modified for clarity and conciseness.

Student 1: I think the width of the pulse on the right in experiment 2 would be greater than it was before. We know that the width of the pulse on the left was doubled in experiment 2, so when the pulse gets to the right spring it should be bigger on the right too.

Student 2: That makes sense to me. Since the pulse on the left had a greater width, the time for it to pass the junction would have been longer in experiment 2. This longer time would cause the pulse on the right to have a greater width than in experiment 1.

Student 3: I don't think the time for the pulse to pass the junction in experiment 2 would be longer. The time for the hand to move to generate the pulse is the same in both experiments, so the bee at the junction should move up and down in the same amount of time in both experiments.

1. With which student(s) do you agree?
2. Explain your response to the previous question.

Figure 3.6: Version 3 analytic support questions.

3.6.3 Version 3 results

The version 3 question sequence was given to students during the Spring of 2021, which was taught remotely due to the COVID-19 pandemic. The quiz was administered between relevant lecture instruction and the corresponding tutorial. A total of 245 students responded.

Responses to the target question on versions 3A and 3B did not differ (chi-square $p = 0.93$) (see Table 3.1). Therefore we conclude that inclusion of the bug questions alone did not influence student thinking on the target question. Additionally, performance on the screening question was not significantly different between versions 3A and 3B (chi-square $p = 0.56$) (See Table 3.2). As a result, we pooled data for the two versions for the rest of the analysis.

Similar to previous versions, asking about consistency between students' responses to the target question and, in this case, the analytic support question did not seem to prompt reflection. Of the 118 students who answered inconsistently, only 31 (26%) recognized their

Table 3.1: Contingency table showing performance on the target question for versions 3A and 3B of the question sequence. Version 3A included bug questions directly before the target question while version 3B presented the same questions after the target question.

	Target correct	Target incorrect	Total
Version 3A	55	114	169
Version 3B	25	51	76
Total	80	165	245

Table 3.2: Contingency table showing screening question performance for versions 3A and 3B of the question sequence. The versions only differed in question order following the screening question.

	Screening correct	Screening incorrect	Total
Version 3A	91	78	169
Version 3B	44	32	76
Total	135	110	245

inconsistency. We considered students to be consistent in each of the following cases:

- They chose "greater than" on the target question, and they agreed with either student 1, student 2, or both students 1 and 2 on the analytic support question,

- They chose "less than" on the target question, and selected "None of the students" on the analytic support question, or

- They chose "equal to" on the target question, and agreed with student 3 on the analytic support question.

All other response combinations were considered inconsistent.

For students who indicated inconsistency in their answers, there was overall improvement on the target question. (Recall that students who indicated consistency were not given the target

Table 3.3: Contingency table showing performance on the target question before and after the intervention for all students who indicated inconsistency in their responses to the analytic support and target questions. Data from versions 3A and 3B are combined.

	Target second response correct	Target second response incorrect	Total
Target first response correct	3	5	8
Target first response incorrect	16	16	32
Total	19	21	40

question again, so they are not included in the following analysis). Table 3.3 shows these students' first and second responses to the target question. A McNemar test with Edward's continuity correction (used when a cell in the contingency table has a value of less than 5) indicates that more students switched from an incorrect answer on the target question to a correct answer after completing the intervention [$p = 0.029$, Cohen's g effect size = 0.26 (large)]. However, of the 165 students who incorrectly answered the first instance of the target question, only 16 indicated inconsistency and went on to change their answer to a correct response, which is a small impact.

3.6.4 Version 4 design

Despite having added several additional questions and the invitation to change their initial answer, the previous interventions in versions 1, 2, and 3 did not lead to a large improvement in performance on the target question. However, we were not prepared to conclude that the overall strategy was flawed. Instead we continued to hone the specific components.

One modification used on version 4 concerned replacing the additional bug questions. Those questions served two purposes. One was to help determine whether the students answering the original screening questions correctly were indeed content-proficient. The other was

In experiment A, a student moves her hand up and down in a time Δt_0 to generate a pulse on the rope shown. A bug of negligible mass rests in the middle of the rope. In experiment B, the student moves her hand up and down with the same amplitude as before, but in half the time: ½ Δt_0.

1. Is the time for the leading edge of the pulse to reach the bug in experiment B *greater than, less than,* or *equal to* what it was in experiment A?
2. Is the time for the bug to move up and then back down again in experiment B *greater than less than,* or *equal to* what it was in experiment A?
3. Explain your responses to questions 1 and 2.

Figure 3.7: Version 4 bug questions, which appeared at the beginning of the sequence before the screening question.

to elicit student knowledge important for the target question, knowledge we speculated was available to students but not highly accessible. Therefore version 4's bug questions dealt with both the time for the pulse to *reach* the junction and the time for it to *pass* the junction. These questions are shown in Fig. 3.7 and appeared before the original screening question. A fuller discussion of these questions can be found in [41].

We also modified the fictional student dialogue (see Fig. 3.8). In this version the two student arguments presented are founded on the idea that the speed of the transmitted pulse is the same in both experiments, affirming a key part of the analysis. However, the fictional students disagree on the time for the pulse to pass the junction. Instead of asking students to agree with one of the fictitious students, they are asked how each of them would answer the target question based on their line of reasoning. The goal of this component was twofold: to expose students to the correct line of reasoning and to induce them to engage in process 2 thinking. Simply presenting the different arguments and asking students which one they agreed with, as in version 3, may have just led them to declare agreement with the one who arrived at their preferred conclusion. We changed the strategy to ask them to complete lines of reasoning

Two students work together to answer [the target question]. The students have the following conversation:

Student 1: The speed of the transmitted pulse should be the same in both experiments because the spring on the right is unchanged. However, since the pulse on the left had a greater width, the time for it to pass the junction would have been longer in experiment 2.

Student 2: I agree that the speed of the transmitted pulse should be the same, but I don't think the time for the pulse to pass the junction in experiment 2 would be longer. This is because the time for the hand to move to generate the pulse is the same in both experiments.

1. Based on the conversation above, how would **student 1** answer [the target question]?
2. Explain your response to the previous question.
3. Based on the conversation above, how would **student 2** answer [the target question]?
4. Explain your response to the previous question.

Figure 3.8: Version 4 analytic support questions.

without regard to their agreement with either the premises or the conclusions [45]. This was meant to avoid the possibility of their simply looking for an argument consistent with their answer and more reliably stimulate engagement of process 2 by requiring students to make deductions based on premises stated by the hypothetical students, analytical reasoning that can be bypassed if a conclusion is known in advance. The questions also required conscious thinking about the motion of the junction - reinforcing knowledge our findings suggest was available to students, even if they had not consciously used it, and foregrounding its relevance.

Finally, we also revised the invitation for students to revisit their initial answer to the target question (see Fig. 3.9) with the hope that it would prompt more students to recognize a need to reflect. In version 4, we asked students if the preceding questions caused them to reconsider their thinking on the target question. It also allowed them the options of

Answering [the analytic support questions] may (or may not) have led you to think more about [the target question] or even reconsider your answer to [the target question]. Below, reflect on how (if at all) your thinking has changed.

a) After answering [the analytic support questions], I would now answer [the target question] differently than before.
b) After answering [the analytic support questions], I would answer [the target question] the same as before, but I would add to or change my explanation somewhat.
c) After answering [the analytic support questions], my answer and explanation on [the target question] are still the same.

Figure 3.9: Version 4 question inviting students to revisit their thinking on the target question.

specifying how their thinking had changed, if at all [46]. The question appeared as depicted in Fig. 3.9 (except with question numbers specified instead of the text in brackets).

3.6.5 Version 4 results

Version 4 was administered on an online quiz in three quarters during which in-person instruction had resumed following the COVID-19 pandemic. In each quarter – Autumn 2021, Winter 2022, and Spring 2022 – the question sequence as before was presented to students after lecture instruction and before the tutorial associated with the online quiz. The sample sizes were 189, 255, and 188 for each respective quarter, and data analysis proceeded as before. The data was combined since it was comparable in that there was no difference in performance on the target question for content-proficient students.

Of the 425 students who incorrectly answered the target question, 280 (66%) indicated that they wanted to change their answer and/or explanation to the target question when asked if they wanted to revisit their thinking, a much larger percentage of students than on version 3.

It seems that this version was more successful at prompting students' desire to revisit their thinking, in line with the intervention strategy component intended to raise doubt and help students avoid reasoning hazard B.

Below, the intervention strategy is assessed using the strategy outlined in Sec 3.5.5. First student improvement on the target question is investigated. Second, this improvement is further analyzed to determine if the intervention is more beneficial for content-proficient students compared to noncontent-proficient students. Third, the data is analyzed to determine if performance on the analytic support questions is associated with improved performance on the target question. Lastly, students' CRT scores are examined in combination with question sequence response data to determine if the intervention is benefiting students regardless of their natural tendency towards cognitive reflection.

Students improve on the target question

Recall that the first level of assessment involved the degree to which the intervention caused students to switch from incorrect to correct answers to the target question. Fig. 3.10 and Table 3.4 show student response patterns for the first and second instances of the target question, which is also reported in [41]. A McNemar test indicates that students were significantly more likely to shift from an incorrect response to a correct one than vice versa [$p = 7.56 \times 10^{-15}$, Cohen's g effect size = 0.34 (large)]. On this version, 112 students improved their answer on the target question from the original 425 students who responded incorrectly to the target question the first time. Recall that on version 3, only 16 out of 165 who gave an initial incorrect response to the target question improved their answer. Therefore, the version 4 intervention reached a larger percentage of students than the version 3 intervention.

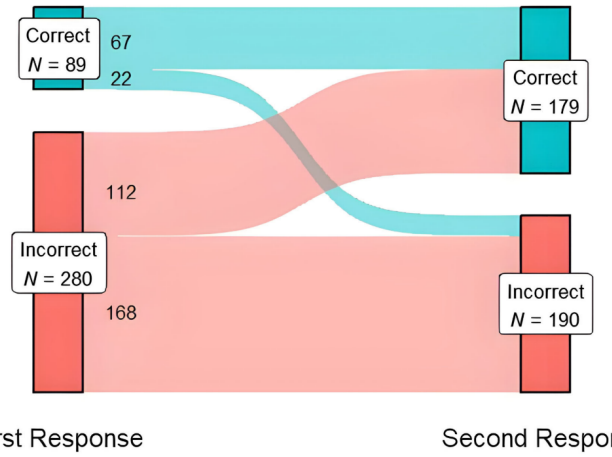


Figure 3.10: Sankey diagram showing student responses to the target question before and after the intervention as presented in [41].

Table 3.4: Contingency table showing performance on the target question before and after the intervention for all students who elected to revisit their initial answer on version 4.

	Target second response correct	Target second response incorrect	Total
Target first response correct	67	22	89
Target first response incorrect	112	168	280
Total	179	190	369

Table 3.5: Contingency table showing performance on the target question after the intervention for all students who elected to revisit their initial answer as a function of their performance on the screening question on version 4.

	Target second response correct	Target second response incorrect	Total
Content proficient	95	46	141
Not content proficient	84	144	228
Total	179	190	369

Content-proficient students improve more than noncontent-proficient students on target

The second level of assessment involved the degree to which improvement on the target question was more pronounced for students who were deemed to be content-proficient by correctly applying relevant content knowledge on the screening question. We conducted a chi-square test comparing the proportions of content-proficient and noncontent-proficient students who ultimately responded with a correct answer on the target question after encountering the intervention (see Table 3.5), which showed that content-proficient students were more likely to respond correctly [$p = 1.18 \times 10^{-8}$, Cramér's V effect size = 0.30 (medium)].

A risk ratio gives more insight into how many times more likely content-proficient students were to correctly reanswer the target question compared to noncontent-proficient students:

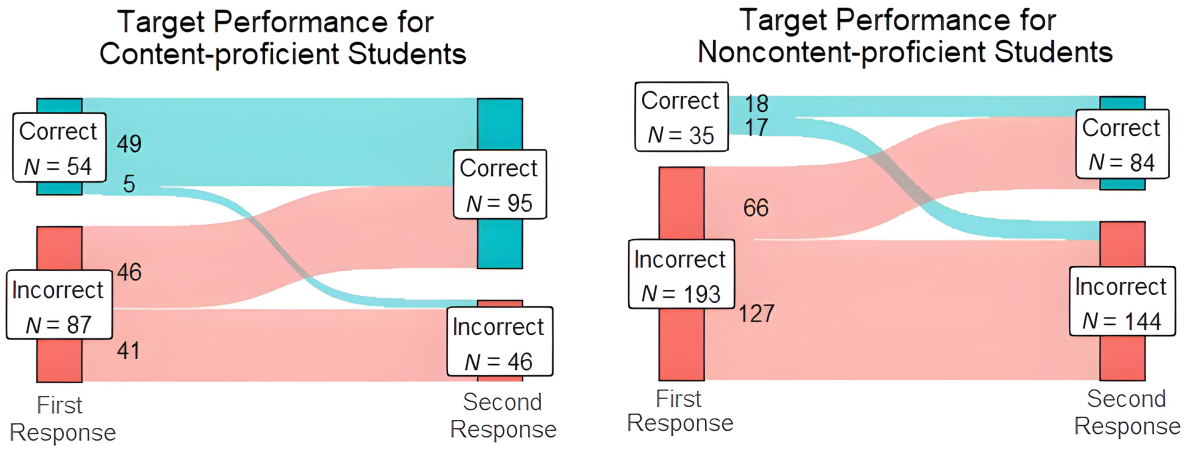
$$\begin{aligned} \text{Risk Ratio} &= \frac{\text{(Target again corr \& content proficient)/Content proficient}}{\text{(Target again corr \& not content proficient)/Not content proficient}} \\ &= \frac{95/141}{84/228} = \frac{0.67}{0.37} = 1.83 \end{aligned} \tag{3.1}$$

We see that content-proficient students were 1.83 times more likely to correctly answer the target question after completing the intervention compared to noncontent-proficient students.

To gain greater insight into student improvement from the first to second instance of the target question, not just performance on the second instance of the target question, we conducted McNemar tests on each group. Sankey diagrams showing answering patterns on the target questions for each pool of students based on their content proficiency are shown in Fig. 3.11, and Tables 3.6 and 3.7 display the corresponding data for content-proficient and noncontent-proficient students respectively. A McNemar test for content-proficient students comparing the number of students who changed their response to the target question from incorrect to correct and vice versa shows a large-effect-size difference, with more students improving accuracy on the target question ($p = 9.40 \times 10^{-9}$, Cohen's g effect size $= 0.40$(large), 95% C.I. [0.30, 0.48]). A McNemar test of the same comparison but for noncontent-proficient students also shows a large-effect-size difference in the same direction ($p = 7.51 \times 10^{-8}$, Cohen's g effect size $= 0.30$(large), 95% C.I. [0.21, 0.38]). However, Cohen's g value for content-proficient students is greater than that for noncontent-proficient students. If these values were significantly different, they would not have overlapping confidence intervals but we do see some slight overlap in the 95% confidence intervals generated through bootstrapping. While there is no statistical difference, there still seems to be a case for content-proficient students having better improvement than noncontent-proficient students.

Similar to above, we ran a chi-square test comparing performance on the latter target question for content-proficient and noncontent proficient students, but only for those who initially answered the target question incorrectly (see Table 3.8). This showed a significant difference that content-proficient students who initially responded incorrectly to the target question were more likely to switch to the correct answer after the intervention than noncontent-proficient students [$p = 0.003$, Cramér's V effect size $= 0.18$ (small)].

The following risk ratio quantifies this difference in likelihood:



(a) (b)

Figure 3.11: Sankey diagrams showing student responses to the target question before and after the intervention on the version 4 question sequence grouped by content proficiency.

Table 3.6: Contingency table showing performance on the target question before and after the intervention for content-proficient students who elected to revisit their initial answer on version 4.

	Target second response correct	Target second response incorrect	Total
Target first response correct	49	5	54
Target first response incorrect	46	41	87
Total	95	46	141

Table 3.7: Contingency table showing performance on the target question before and after the intervention for noncontent-proficient students who elected to revisit their initial answer on version 4.

	Target second response correct	Target second response incorrect	Total
Target first response correct	18	17	35
Target first response incorrect	66	127	193
Total	84	144	228

Table 3.8: Contingency table showing performance on the target question after the intervention for students who initially answered the target question incorrectly and elected to revisit their initial answer as a function of their performance on the screening question on version 4.

	Target second response correct	Target second response incorrect	Total
Content proficient	46	41	87
Not content proficient	66	127	193
Total	112	168	280

$$\begin{aligned}\text{Risk Ratio} &= \frac{\text{(Target again corr \& content proficient)/Content proficient}}{\text{(Target again corr \& not content proficient)/Not content proficient}} \\ &= \frac{46/87}{66/193} = \frac{0.53}{0.34} = 1.55\end{aligned} \tag{3.2}$$

For students who incorrectly answered the target question, content-proficient students were 1.55 times more likely to correctly answer the target question after completing the intervention compared to noncontent-proficient students.

It would appear that content-proficient students tend to improve on the target question to a greater degree than noncontent-proficient students, but only weakly. It could be that the screening question does not require all the mindware necessary for correctly answering the target question or it may not require the level of development in formal knowledge needed to aid in the detection and override of incorrect intuitive ideas on the target question. In this case, some students we are identifying as content-proficient may be answering the target question incorrectly due to insufficient mindware as opposed to insufficient reasoning. These students would be less likely to benefit from the intervention, and so the group of students deemed "content-proficient" would not have as drastic of an improvement on the target question compared to those labeled "noncontent-proficient."

The weak relationship we are seeing here could also be a result of the mechanism by which the intervention causes improvement. If the intervention is helping students develop understanding of content knowledge in addition to prompting process 2 engagement, then students who were considered noncontent-proficient before the intervention would now have the conceptual understanding necessary to succeed on the target question.

Overall, the intervention appears successful at encouraging students to revisit their thinking and improve on the target question. There is a slight advantage for content-proficient students as expected, but the results could be impacted by other factors that are weakening the observed effect of this difference.

Table 3.9: Contingency table showing the decision to revisit the target question after the intervention as a function of student performance on the analytic support questions on version 4.

	Wanted to revisit	Did not want to revisit	Total
Analytic support correct	183	146	329
Analytic support incorrect	186	114	300
Total	369	260	629

Analytic support question performance is not associated with revisiting or improvement on target

The third level of assessment concerned the impact, if any, of answering the analytic support questions (related to the fictitious student dialogue) correctly. Table 3.9 shows whether or not students chose to revisit their thinking on the target question based on their performance on the analytic support questions. Using a chi-square test, there is no difference in the desire to revisit ($p = 0.105$).

We also conducted a chi-square test to see if performance on the target question differed depending on analytic support question performance for those who chose to revisit the target question (see Table 3.10). We see that there is no significant difference between these two groups ($p = 0.132$).

One might have expected that correctly answering the analytic support questions would provide an advantage with regards to these students' decision to revisit their thinking and improved target performance since they successfully considered the relevance of the hand motion on the pulse width. However, students who answered the analytic support questions incorrectly were just as likely to revisit and improve performance on the target question.

Accuracy on the analytic support questions depended on correctly determining the con-

Table 3.10: Contingency table showing performance on the post-intervention target question for students who elected to revisit their initial answer as a function of their performance on the analytic support questions on version 4.

	Target second response correct	Target second response incorrect	Total
Analytic support correct	96	87	183
Analytic support incorrect	83	103	186
Total	179	190	369

clusions that fictional student 1 *and* student 2 would make on the target question. We considered the possibility that correctly following the line of reasoning for student 2, who had correct ideas, might be sufficient for successfully arriving at the correct conclusion on the post-intervention target question. Table 3.11 shows student performance on the second instance of the target question for those who opted to revisit their thinking based on their performance following student 2's reasoning only. Students who correctly followed student 2's reasoning and chose to revisit the target question were more likely to reanswer the target question correctly [$p = 0.004$, Cramér's V effect size $= 0.15$ (small)] than those who did not select a correct conclusion for student 2.

The following risk ratio indicates that students who determined the correct conclusion for student 2 were 1.50 times more likely to correctly respond to the target question than those who selected an incorrect conclusion for student 2.

$$\begin{aligned} \text{Risk Ratio} &= \frac{\text{(Target again corr \& Student 2 response corr)/Student 2 response corr}}{\text{(Target again corr \& Student 2 response incorr)/Student 2 response incorr}} \\ &= \frac{148/281}{31/88} = \frac{0.53}{0.35} = 1.50 \end{aligned} \tag{3.3}$$

It appears that successful student engagement with the correct line of reasoning was pro-

Table 3.11: Contingency table showing performance on the target question after the intervention for students who elected to revisit their initial answer as a function of their performance on the analytic support question asking how student 2 would respond to the target question on version 4.

	Target second response correct	Target second response incorrect	Total
Analytic support student 2 correct	148	133	281
Analytic support student 2 incorrect	31	57	88
Total	179	190	369

ductive in helping students improve on the target question for those who elected to revisit their thinking. This suggests that the analytic support question regarding the correct line of reasoning is the catalyst for student improvement. However, this is only supported with a small effect size difference.

It is possible that improvement on the target question showed no dependence on performance on the analytic support questions as a whole because they were able to prompt process 2 engagement such that even if reasoning wasn't applied correctly to the analytic support questions, analytic thinking could still be applied correctly when responding to the post-intervention target question. In any case, version 4 was able to improve student performance on the target question with greater impact than previous versions, and so it appears that the analytic support questions and opportunity to revisit as presented on this version were responsible even if successfully following the reasoning of both fictitious students was not necessary for said improvement.

Table 3.12: Contingency table showing content-proficient students' decision to revisit their thinking on the target question based on CRT score level for version 4.

	Wanted to revisit	Did not want to revisit	Total
Low CRT (0-1)	38	36	74
High CRT (2-3)	103	106	209
Total	141	142	283

Content-proficient student improvement does not depend on CRT score

The fourth level of assessment concerns the degree to which content-proficient students with higher CRT scores benefit from the intervention compared to content-proficient students with lower CRT scores. First, a chi-square test showed that content-proficient students with differing CRT scores had an equally-likely desire to revisit the target question (see Table 3.12) ($p = 0.76$). In other words, the invitation to revisit seemed to prompt students to initiate the process of sustained effort regardless of their natural propensity towards cognitive reflection.

Beyond a mere decision to revisit thinking, the following analyses investigate if performance on the second instance of the target question was different for students with differing CRT scores. Table 3.13 shows content-proficient students' performance on the second instance of the target question based on CRT score level. A chi-square test showed no difference in performance for these two groups ($p = 0.292$). Additionally, when making the same comparison for only those students who originally answered the target question incorrectly (see Table 3.14), there was no difference in the fraction of those who changed to the correct answer on the target question based on their CRT score level ($p = 0.412$). This suggests that the intervention was able to support process 2 thinking in arriving at a correct conclusion equally for students who would naturally tend to engage in process 2 thinking as well as those who would not typically do so.

Lastly, we compared McNemar tests (with Edward's continuity correction) for content-

Table 3.13: Contingency table showing performance on the target question after the intervention for content-proficient students who elected to revisit their initial answer as a function of their performance on the CRT for version 4.

	Target second response correct	Target second response incorrect	Total
Low CRT (0-1)	23	15	38
High CRT (2-3)	72	31	103
Total	95	46	141

Table 3.14: Contingency table for students who answered the target question incorrectly showing performance on the target question after the intervention for content-proficient students who elected to revisit their initial answer as a function of their performance on the CRT for version 4.

	Target second response correct	Target second response incorrect	Total
Low CRT (0-1)	12	14	26
High CRT (2-3)	34	27	61
Total	46	41	87

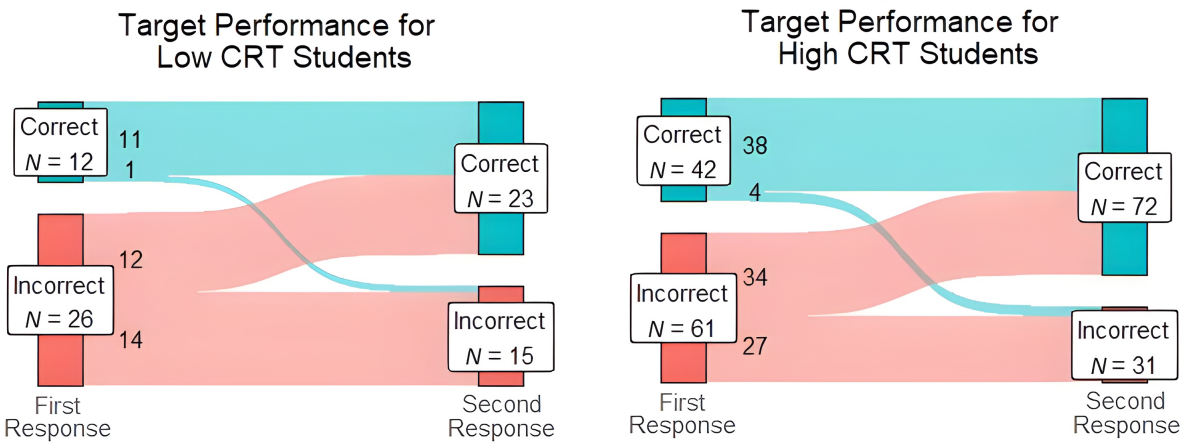


Figure 3.12: Sankey diagrams showing version 4 content-proficient student responses to the target question before and after the intervention for students with low or high CRT scores.

proficient students with low and high CRT scores to see if there was a difference in the degree to which accuracy increased on the target question as opposed to decreased. Fig. 3.12 show Sankey diagrams of content-proficient students' response patterns on the target question based on their CRT score, and Tables 3.15 and 3.16 tabulate the same information. For those with low CRT scores, there was a large effect size showing improvement ($p = 0.005$, Cohen's g effect size = 0.42 (large), 95% C.I. [0.25, 0.50]). For those with high CRT scores, there was also a large effect size showing improvement ($p = 2.53 \times 10^{-6}$, Cohen's g effect size = 0.40 (large), 95% C.I. [0.29, 0.48]). These effect sizes appear to be no different since their confidence intervals completely overlap.

Note that preintervention target performance for this data set was associated with CRT score for content-proficient students [41]. Those with high CRT scores were more likely to answer the target question correctly and CRT score was a predictor of target performance. This indicated that responding to the target question was largely a matter of engaging in cognitive reflection on an unproductive process 1 model.

Table 3.15: Contingency table showing performance on the target question before and after the intervention for content-proficient students with low CRT scores who elected to revisit their initial answer on version 4.

	Target second response correct	Target second response incorrect	Total
Target first response correct	11	1	12
Target first response incorrect	12	14	26
Total	23	15	38

Table 3.16: Contingency table showing performance on the target question before and after the intervention for content-proficient students with high CRT scores who elected to revisit their initial answer on version 4.

	Target second response correct	Target second response incorrect	Total
Target first response correct	38	4	42
Target first response incorrect	34	27	61
Total	72	31	103

However, this trend may not be expected on the target question following the intervention. In Mays *et al.* [47], the authors implemented a similar intervention strategy where students explored alternative lines of reasoning. They stated that one could argue that a student's tendency towards cognitive reflection may not be expected to correlate with performance on their post-intervention target question since the intervention task explicitly had students analytically engage with different lines of reasoning. The same could be anticipated for our intervention since the analytic support question encourages students to follow different reasoning paths for each of the fictitious students' arguments. Students who engage with the analytic support question in this way have already utilized process 2 with regards to the target question, and so their natural propensity to engage in this kind of thinking may be less of a factor when reanswering the target question themselves.

3.7 Discussion of Limitations and Further Research

It is important to note that our data come from a single population consisting of undergraduate physics students at the University of Washington. Also, many of our analyses focused on content-proficient students, who comprised anywhere from 44 to 140 students during a single quarter in our study. This meant that some of our samples were relatively small.

The intervention we designed was confined to a written question sequence about a pulse on a spring system and tested for effectiveness immediately thereafter. Therefore, our intervention may not have a long-term positive effect on improving students' overall reasoning in physics. However, it is possible that the use of similar interventions used consistently throughout a physics course could prompt students to be more reflective reasoners, though research would need to be done to evaluate this.

Other studies [27] have reported on interventions that provide more benefit to students with higher CRT scores. It is not clear why the same isn't true here, although we suspect it could be due to the nature of our analytic support questions and timing of the post-intervention target question. In [27], their final instance of the target question was presented on an exam

some time after their intervention was implemented. At that point, it is not unexpected that students with a higher propensity towards cognitive reflection would perform better than those with more of a propensity towards miserly thinking. Since students in this study were given the target question again directly after encountering our intervention, the intervention's encouragement to engage in process 2 thinking may have caused students' natural tendency towards cognitive reflection to be less relevant in this context. The nature of the analytic thinking questions to prompt students to follow alternative reasoning paths (similar to [47] may have caused students to reflect and ultimately reason on the path of sustained effort when they wouldn't have otherwise, making an association between CRT score and improved performance on the target question unlikely.

There are also nuances to students' CRT scores that could be impacting the results. We note that the pool of students in this study tends to already have quite high CRT scores and that there were some unexplained variations in results from version to version. We also note that we used the 3-item CRT; however, a 7-item CRT also exists which may do better at discriminating between students' cognitive reflection skills. This more extensive CRT will be used and analyzed in relation to question sequences involving physics content that includes the pulse questions in addition to other topics. There is also research being conducted to determine if the CRT is an independent measure of cognitive reflection or if it is associated with other academic measures such as course grades.

Beyond seeking to understand students' reasoning processes in physics, the ability to design DPTs-based written intervention question sequences in multiple contexts could be instructionally useful for encouraging students to use cognitive reflection in a productive way. Developing similar interventions in other content areas of physics would provide a proof of concept for using dual-process theories to help students in their reasoning processes. Data from a collection of interventions could be analyzed to identify mechanisms that help students reason for the purpose of informing curriculum development.

3.8 Conclusion

This study aimed to investigate student reasoning in introductory physics while developing a question sequence with an intervention designed to address students' reasoning inconsistencies, specifically in the context of wave dynamics.

We presented introductory physics students with a question sequence designed to identify students making reasoning errors as described by DPTs and to encourage effective process 2 engagement through a written intervention. Students who responded correctly to the screening question were considered to be content-proficient, yet on all versions combined, only 47% of these students correctly responded to the target question [41], consistent with the idea that they relied on an unproductive process 1 response and did not effectively engage in process 2 to arrive at the correct answer.

Through an iterative process, we tested various interventions consisting of combinations of additional questions, analytic support, and reflection prompts. The ineffectiveness of the first few versions of the intervention indicates that simply drawing students' attention to the vertical motion of the spring, and giving them an opportunity to change their responses, is not sufficient to increase performance on the target question. Evidently, inducing reflection and supporting analysis is more difficult than we initially expected. From the perspective of DPTs, this result suggests that the appeal of the incorrect process 1 result can be very strong. It may even be the case that slowing down and thinking may cause the intuitive response to become more firmly held, in a form of confirmation bias. These results also demonstrate that the mere presence of related questions and the opportunity to revisit are insufficient for generating more correct answers. Thus we show that turning a broad strategy with theoretical support into a specific intervention is difficult.

By iteratively redesigning the intervention we arrived at a version that succeeded in improving performance on the target question for a higher percentage of students. We believe that the intervention succeeded because it raised sufficient doubt in students' process-1-based re-

sponses, reminded students of relevant mindware, and prompted them to consider alternative lines of reasoning that proved necessary to engage process 2 in a productive way.

The intervention was successful in causing a significant shift in answers from incorrect to correct, especially for students who demonstrated content proficiency by correctly responding to the screening question. While it is possible that the intervention may have further developed students' understanding about the motion of the junction, if students were performing poorly on the target question due to inadequate conceptual understanding alone, this intervention would not have resulted in the impact seen. Since performance on the target question was also dependent on cognitive reflection, the intervention seemed to target student reasoning processes, lending credence to the DPTs perspective.

It is important to discuss that our DPTs-based intervention was targeted at content-proficient students who were more likely to tend towards cognitive frugality (measured by a low score on the CRT). We note that while content-proficient students' propensity towards cognitive reflection, as measured by the cognitive reflection test, was related to their initial response to the target question, it was not related to their subsequent response to the target question after encountering our intervention. However, this could be due to the intervention supporting reflection and process 2 thinking for all students regardless of their natural tendency towards miserly reasoning. We also note that there is only weak evidence to support that content-proficient students were benefiting to a greater degree from the intervention than noncontent-proficient students as expected.

Other studies have used DPTs to guide improvement in student reasoning using different methods such as additional instruction, in-class activities, group work, multiple choice homework questions, and reasoning chain construction tasks [24, 27, 46, 47, 26, 12, 48]. In particular, Lindsey *et al.* [28] showed that a strategy based on getting students to engage with their incorrect reasoning, attributed to faulty process 1 output and supported by heuristics that are tightly linked to content that is explicitly taught, but not universally applicable, can be effective, and in fact more effective than an intervention based on bolstering the correct ideas

that underlie the appropriate analysis and bypassing incorrect thinking. Our intervention is similarly based on engagement with the incorrect ideas and also showed improvement. Given the many differences, such as assessment schemes, our results can't be directly compared to theirs. Instead our results can be seen as supporting the merits of the general approach, even in scenarios that differ substantially in the patterns of reasoning displayed.

It is apparent that the appropriate use of cognitive reflection, the propensity to mediate intuition with analytic thinking, is an important skill needed for succeeding in physics beyond content understanding. Therefore, physics instruction should attend to cognitive reflection in addition to conceptual understanding. Although it has yet to be shown, it is possible that the cumulative effect of interventions designed to induce and support process 2 thinking may have a longer-term or more general impact. The question sequence we developed provides an opportunity for students to practice cognitive reflection in a way that does not overburden the course and could be replicated in other content areas.

3.9 Acknowledgments

The authors appreciate the many contributions of Andrew Boudreaux, Mila Kryjevskaia, Beth Lindsey, Drew Rosen, and MacKenzie Stetzer to this study. We also appreciate the cooperation of the many instructors who allowed us to modify course materials and collect student response data in their courses. Special recognition goes to Peter Shaffer, Nikolai Tolich, Kazumi Tolich, and David Smith for their willingness to allow us access to the relevant instructional platforms, granting us the ability to modify online quizzes and collect data. We would also like to thank the members of the University of Washington Physics Education Group for their input. This material was 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.

Chapter 4

NEWTON'S SECOND LAW QUESTION SEQUENCE TO EVALUATE AND ADDRESS STUDENT REASONING INCONSISTENCIES: AN APPLICATION OF DUAL-PROCESS THEORIES

4.1 Abstract

Human reasoning can be characterized by dual-process theories, which suggests that there are two processes involved in thinking and decision-making. Process 1 automatically generates a mental model to address the particular situation at hand, while process 2 may or may not engage to assess the provided model. Student inconsistencies in response to physics questions can stem from reasoning that, in some cases, relies on incorrect process 1 ideas. Previous research by Kellar and Heron has investigated question sequences in the context of pulses on a spring to (1) distinguish between students' conceptual understanding and reasoning approaches and (2) develop an intervention to help support students' effective process 2 engagement. Here, a similar question sequence in the context of forces and Newton's laws is investigated. The sequence appears to achieve the two main desired purposes described above consistent with the previous research in a different context, and serves as another example of how an understanding of reasoning mechanisms can be used to support students' cognitive reflection on tasks that tend to elicit incorrect intuitive ideas.

4.2 Introduction

An intriguing phenomenon has been reported in physics education that students, when tasked with seemingly isomorphic questions, may answer one correctly and one incorrectly even if

seen one after the other [24, 29, 27, 22]. While they may have demonstrated appropriate content knowledge on one question, they may fail to apply it to the other question. This error may be due to insufficient reasoning.

The perspective of dual-process theories (DPTs) has been used to characterize this type of answering pattern in terms of the interplay between two types of reasoning processes [19]. Process 1 is intuitive and automatic while process 2 is analytic and deliberate. In any given situation, one's process 1 always activates to generate an initial idea to address the situation. Process 2 may or may not be engaged to evaluate the plausibility of this initial idea. If process 1 offers an incorrect intuition and process 2 is not effectively engaged to override this notion, then an incorrect conclusion can be drawn.

A variety of studies have been conducted to evaluate the explanatory power of DPTs with regard to student answering patterns in physics [24, 26, 12], and several more have leveraged understanding of DPTs to modify instruction intended to improve student reasoning, either by strengthening correct process 1 ideas or encouraging effective process 2 engagement [29, 26, 27, 49, 28]. Kellar and Heron [41, 30] have previously contributed to this effort using a question sequence in the context of a pulse on a spring. The pulse sequence started with "screening" and "target" questions adapted from [24] designed to identify students who possessed adequate conceptual understanding yet relied on an incorrect process 1 model. In [41], the authors developed a method to test DPTs-based assumptions surrounding the screening and target questions. Following the screening and target questions, the sequence then included a "micro-intervention" designed to raise doubt in faulty intuitive ideas, cue relevant knowledge, and support process 2 reasoning. Students were prompted to contemplate alternative lines of reasoning and consider revisiting their own thinking. This efficacy of this intervention strategy was evaluated in [30].

This investigation extends the prior work on the pulse question sequence by adapting it to a new context: forces and Newton's laws. The screening and target questions used here involve boxes at rest on rough surfaces (see Fig. 4.4), which have been shown to exhibit

student answering patterns characteristic of dual-process theories of reasoning [29, 12, 49]. This study provides evidence for the generalizability of the methods used in [41] and [30] and adds to the collection of DPTs-based interventions in physics.

This chapter begins by providing an overview of dual-process theories as it relates to the works cited above. It then describes the investigation including the instructional context and question sequence. Results are described in detail and evaluated in view of DPTs and existing research, including a comparison of results in the pulse and forces contexts. Finally, discussions of limitations, avenues for further research, and instructional implications are included.

4.3 Theoretical Framework

This work is founded on an understanding of human reasoning outlined by the framework of dual-process theories. There are many different theories within this framework, one of which is described by [22] and visualized in Fig. 4.1. This is the theory that underlies this work. As mentioned above, DPTs suggest there are two main processes involved in reasoning and decision-making: process 1 is heuristic and automatic while process 2 is analytic and deliberate [19]. Upon encountering a task, process 1 engages to offer a provisional model, informed by intuition, prior knowledge, and contextual cues, that serves as an entry point for addressing the task. This process happens automatically and subconsciously and can be thought of as one's "gut reaction." At this point, the reasoner could accept this initial model as grounds for a conclusion and end the entire reasoning operation related to the task at hand without engaging process 2. This is considered the *path of cognitive frugality* [22].

On the other hand, after process 1 generates an idea, process 2 may work to assess that idea during what is considered the *path of sustained effort* [22]. Process 2 (in conjunction with process 1) can serve two main capacities: to seek supporting evidence to justify the provisional model generated by process 1, or to test the model by attempting to disprove it [23]. Either way, after process 2 has deliberately evaluated the initial model, it can either

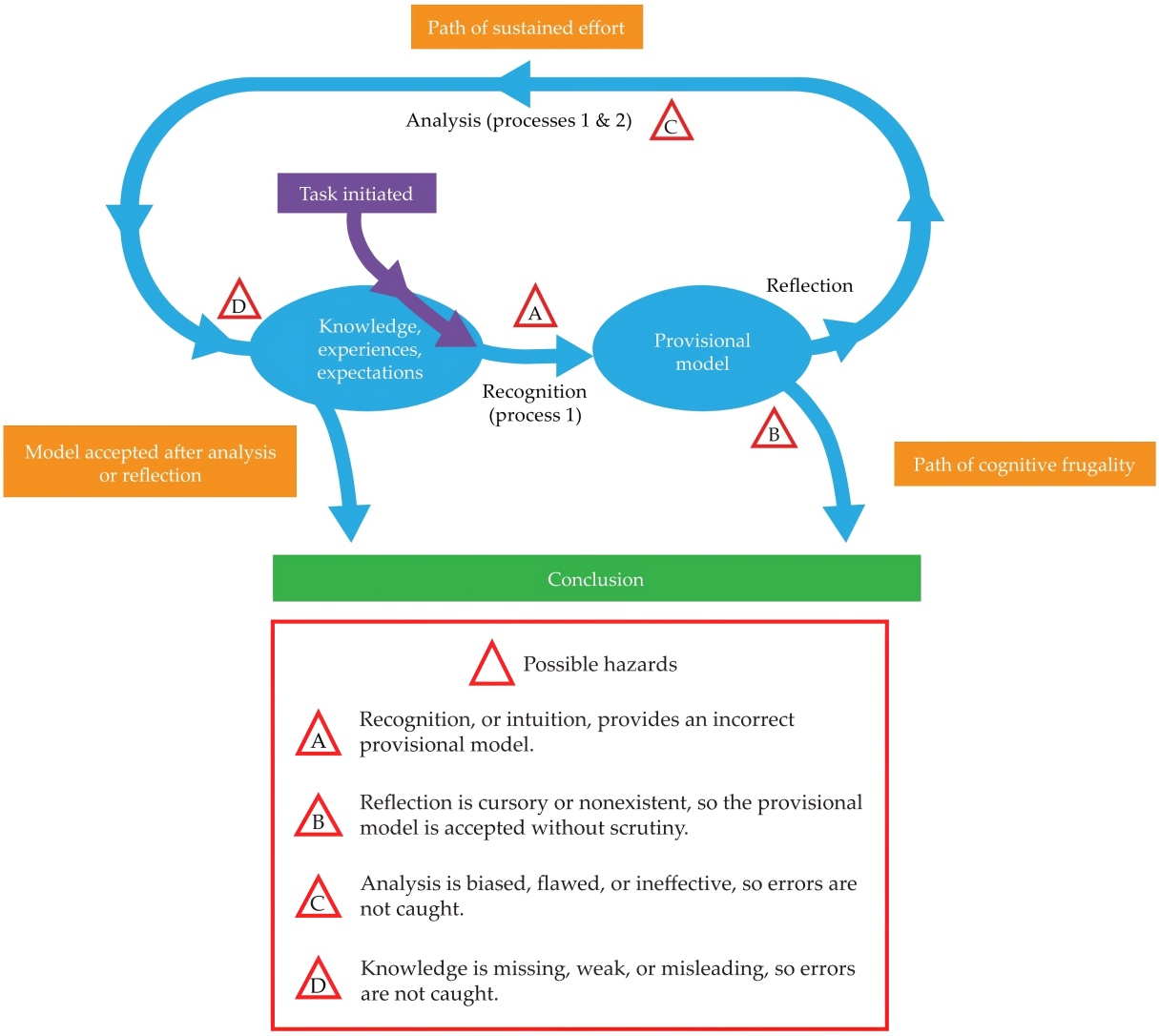


Figure 4.1: Reasoning pathways and hazards from [22].

output a conclusion based on this model or decide that a new model is needed. If a new model is deemed necessary, process 1 is activated again and the reasoning cycle repeats.

The effectiveness of the entire reasoning operation in arriving at a satisfactory conclusion depends on several factors, and there are multiple hazards that can hinder successful reasoning. Broadly, one must have the appropriate content knowledge and skills relevant to the situation, also known as "mindware" [20], in order to succeed. Without mindware, process 1 cannot generate a correct provisional model due to a lack of relevant experiences or expectations.

However, even equipped with the necessary mindware, one may still fail to produce a productive initial idea potentially due to distracting contextual cues related to the task. Providing an incorrect provisional model is denoted by reasoning hazard A in Fig. 4.1. In the case where an incorrect model is generated, the reasoner must be able to override this intuitive thought with a correct conclusion in order to succeed. In this case, "cognitive reflection" is a necessary component of effective reasoning. Cognitive reflection is defined by Frederick as "the ability or disposition to resist reporting the response that first comes to mind" [21], and more highly developed cognitive reflection skills can help reasoners recognize the need to engage in process 2 thinking. If one generates an incorrect model during process 1 thinking and fails to recognize the need to reflect through the engagement of process 2, this is considered falling into reasoning pitfall B.

Unfortunately, mere activation of process 2 is not enough to guarantee a correct conclusion. In the case of an incorrect provisional model, process 2 must perform hypothesis testing strategies such as actively seeking alternative solutions or generating counterarguments. If instead, process 2 performs biased reasoning that rationalizes the provisional model [23], the reasoner falls prey to hazard C.

Finally, in the case where an incorrect intuition is generated by process 1, and the reasoner recognizes the need to reflect and engages in process 2 thinking that uses hypothesis testing

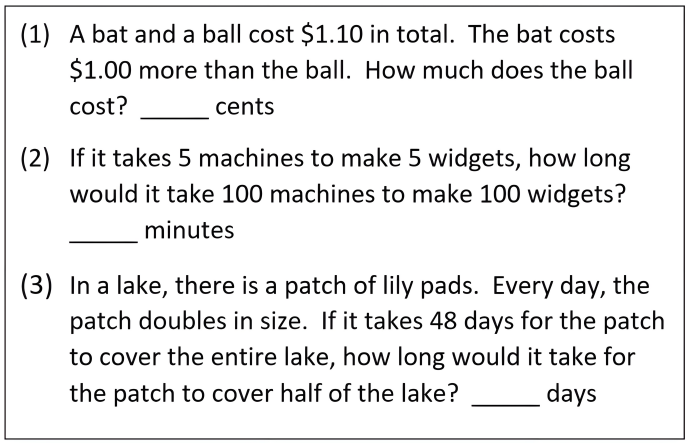


Figure 4.2: The three-item cognitive reflection test developed by Frederick [21].

strategies, one must also have a sufficiently robust conceptual framework to ultimately succeed in overriding the initial model with a correct conclusion. For example, in the presence of competing information within the task, reasoners must be able to distinguish which factors are irrelevant or less important compared to others, leading to the detection of errors and successfully overriding them. Weak mindware that results in the failure to detect errors is denoted by reasoning hazard D in Fig. 4.1.

Now that the various factors impacting human reasoning within the DPTs framework have been outlined, let us return briefly to the concept of cognitive reflection. This tendency is often measured using the three-item Cognitive Reflection Test (CRT) [21] (see Fig. 4.2). A seven-item version has also been developed by [50], which includes each of the three-item CRT questions. The questions consist of word problems requiring simple mathematical applications that tend to generate incorrect intuitive answers. As such, one must engage in minimal reflection to arrive at the correct answer. Each item is scored for correctness, earning either 0 or 1 point. The CRT, in both the three-item and seven-item formats, has been shown to be a powerful predictor of performance on various rational thinking tasks and a reliable measure of the tendency toward miserly processing [37, 50].

Given all of the factors influencing effective reasoning within DPTs, one may wonder how reasoning can be improved, especially as it relates to students. There are two main strategies that can be used to this end: (1) providing practice that automates a direct path to the solution, increasing the likelihood that process 1 generates a correct model, and (2) supporting effective process 2 engagement [22]. Each approach has its advantages, and both can be helpful for improving reasoning outcomes in various contexts.

4.4 Background and Motivation

This work draws on prior dual-process theories investigations of a question sequence in the context of a pulse on a spring [24, 41, 30]. In [24], Kryjevskaia *et al.* developed "screening" and "target" questions designed to gain insight into student reasoning. The screening question(s) involved straightforward assessment of certain physics content. The target question was analogous in the sense that it was intended to require the same content knowledge; however it contained a *salient distracting feature* [14, 32, 33] prompting incorrect process 1 model generation. It is thus hypothesized that success on the target question largely depends on a student's use of cognitive reflection.

Performance on the screening question was used to identify students with relevant content knowledge. Subsequently, the target question was used to observe student reasoning behavior: students who demonstrated conceptual understanding on the screening question but answered the target question incorrectly were believed to have relied on process 1 intuition, consistent with DPTs.

In [41], Kellar and Heron tested this interpretation in the case of the pulse-on-a-spring screening and target sequence provided by [24]. In particular, two main assumptions were assessed, each with two implications:

Assumption (1) The screening question accurately identifies students as having, or not having, the conceptual understanding needed to correctly answer the target question.

Implication (i) A correct answer to the target question depends on a correct response to the screening question.

Implication (ii) Students who correctly answer the screening question should be able to correctly answer related questions that lack salient distracting features.

Assumption (2) For students who demonstrate content understanding by responding correctly to the screening question ("content-proficient students"), responding to the target question is largely a matter of cognitive reflection on incorrect process 1 ideas.

Implication (iii) Content-proficient students who have a greater tendency for cognitive reflection should be more likely to respond correctly to the target question.

Implication (iv) Interventions intended to support student reflection, without providing instruction on relevant content, should increase student performance on the target question.

In [30], Kellar and Heron evaluated the effectiveness of a DPTs-based intervention sequence intended to improve student performance on the pulse target question. In general, there are two possible approaches for leveraging dual-process theories to improve student performance on questions that tend to elicit an incorrect intuitive model as mentioned in Sec. 4.3. One strategy is to automate a correct process 1 response in a given context (reflexive). The other strategy is to support process 2 engagement such that students can override an initial incorrect idea by effectively reasoning on the path of sustained effort (reflective) [22]. The intervention in [30] used the latter strategy.

Research suggests that in order for students to abandon an incorrect intuitive model, they must have some doubt regarding their initial model and they must have the content knowledge necessary to produce the correct model [42, 20, 43]. Given these criteria, an effective reflective DPTs-based intervention that is intended to address student reasoning on a target

question should encompass three main principles:

Principle (1) For students experiencing reasoning hazard A on the target question, they must experience doubt in their process 1 model in order to recognize the need to reflect.

Principle (2) Relevant content knowledge and skills needed to successfully navigate the path of sustained effort must be readily accessible.

Principle (3) Students must have sufficient motivation to engage in productive analytic thinking on the target question in order to produce a correct model.

These principles were the basis for the intervention method described in [30], which consisted of a series of questions related to a fictitious student conversation regarding the target question. The student discussion included information that was intended to raise doubt and remind students of relevant mindware, addressing design principles (1) and (2). Students were then asked how each of the fictitious students would respond to the target question, a task that required consideration of different lines of reasoning in accordance with design principle (3).

After considering the responses of the fictitious students, the next question in the intervention sequence asked students to reflect on how (if at all) their thinking had changed on the target question. Then all students were given the target question again.

In this chapter, a different screening-target sequence involving boxes at rest on rough surfaces (box-friction questions) described in the literature [29, 12, 49] (see Fig. 4.4) is investigated as well as an intervention with similar structure to [30]. The goal is to determine if the previous results are generalizable and to add to the body of DPTs research in physics education.

4.5 Prior Research

Multiple studies have investigated the box-friction questions through the lens of DPTs and tested various intervention strategies, both reflective and reflexive. A summary of the liter-

ature related to this context is presented below.

Kryjevskaia *et al.* [29] designed and implemented metacognitive interventions under the reflective approach. In one intervention, students received metacognitive questions after seeing the box-friction screening and target questions on an exam. The metacognitive questions asked students what an intuitive answer would be to the target question and if they used intuitive thinking. Additionally, students partook in group interviews where they worked in pairs on the screening and target questions. Neither intervention had an effect on student reasoning, and the authors found that incorrect intuitive reasoning approaches were persistent among students.

In a subsequent study conducted by Speirs *et al.* [12], the researchers tested DPTs interpretations using multiple experiments and multiple screening-target pairs including the box-friction questions. The box-friction target question was given as a *reasoning chain construction task*, which allowed students to respond by arranging provided statements related to conceptual knowledge and scenario observations into a logical progression of ideas leading to a conclusion. Students were informed that all of the statements were true and asked to construct a solution to the problem by selecting statements from the provided list, arranging them, and using provided connecting words like "and" and "because" as necessary.

Students were given two different versions of the task, where the treatment group received an additional reasoning element that the control group did not see. This additional statement, called the *analytic intervention element* (AIE), was as follows: "the coefficients of friction are not relevant to this problem." The purpose of this element was to refute the common incorrect intuitive idea associated with the coefficients of friction. Results showed that students who received this statement were more likely to abandon this sort of intuitive reasoning than those who did not have this reasoning element. The AIE seemed to raise doubt in incorrect process 1 thinking, providing a reflective intervention approach.

Lastly, in [49], Speirs *et al.* compared the effectiveness of reflexive and reflective interventions

in relation to the same box-friction questions. In the reflective approach, students were given three tasks similar to the screening question but with various levels of extra, irrelevant variables. After responding to the three tasks, students were asked to reflect on their reasoning for the whole set and asked why the extraneous information was not needed to answer the questions. They were then given solutions, including information about why the extraneous information was not needed and that Newton's second law should be used instead. This intervention was intended to guide students towards engaging process 2 thinking by having them reflect on the application of Newton's second law in the context of static friction.

In the reflexive approach, students were given 30 multiple-choice practice questions. Upon answering each question, students were given immediate feedback as to whether their response was correct or incorrect. The activity aimed to draw students' attention away from the distracting coefficient of friction feature and focus it instead on the relevant forward force feature through repeated practice. In this way, the reflexive intervention intended to promote correct process 1 thinking.

Both methods improved student performance, but the reflexive approach resulted in a larger improvement in this context. The authors concluded that a reflective approach seems to be less effective in a context with salient distracting features, and that better instruction would include interventions targeted at process 1 thinking.

4.6 Investigation

This study seeks to evaluate a question sequence in the context of forces and Newton's laws that contains the box-friction screening and target questions used in [29, 12, 49] as well as an intervention sequence designed according to the methods in [30]. In particular, this sequence serves as a proof of concept for DPT's-based validation methods surrounding screening-target pairs and interventions aimed at supporting process 2 thinking [41, 30].

Screening-target validation methods outlined in [41] are used to test how well the screening question is identifying students with relevant content knowledge and if responding to the

target question is largely a matter of overriding incorrect intuitive ideas. The intervention is evaluated for its alignment with the design principles outlined in Sec. 4.4, namely that it promotes recognition of the need to reflect on intuitive thinking and motivates analytical thinking resulting in production of the correct conclusion on the target question. Students' propensity towards cognitive reflection is evaluated in tandem with their responses to the question sequence.

4.6.1 Context

This investigation was conducted in the context of an introductory calculus-based mechanics course at the University of Washington (UW). The course included lecture, lab, and tutorial components. The data primarily come from three academic quarters starting Autumn of 2022 and ending Autumn of 2024 (see Sec. 4.7.1 for a more in-depth discussion of the data).

The question sequences administered in this study were presented to students on online quizzes called "tutorial pretests," which preceded small-enrollment course sections where students worked in groups on *Tutorials in Introductory Physics* [51]. Tutorial pretests were given most weeks as part of the existing course structure, and so the sequences did not present a significant disruption to the class. The pretests were typically available for two and a half days over a weekend and had a 15-minute time limit, which was intended to be enough time for students to complete the questions without being rushed. Data obtained during Autumn of 2022 and the first administration of the questions during Spring of 2024 followed the usual time limit since the question sequences given were already part of the quizzes used in the course. However, for all other data, there was no time limit imposed since the question sequence included additional questions as part of an intervention. Students received completion scores for attempting each quiz regardless of accuracy, and these scores only comprised a small percentage of students' overall grade in the course.

The seven-item CRT was administered on a separate tutorial pretest in each of several academic quarters in each introductory calculus-based course. This meant that some students

took the test more than once, in which case their first CRT score was used for analysis.

Only students who did not opt out of participation in the study were included as part of the investigation.

4.6.2 Question sequence

In most cases, the question sequence used in this study consisted of a series of questions shown in Fig. 4.3. The screening and target question appeared on a single page in the online pretest, followed by a separate page containing an intervention comprised of analytic support questions and a question asking students if they wanted to revisit their response to the target question. Finally, all students were given the target question again on another page. Students were able to go back to previous pages; however, on the revisiting question, students were instructed not to go back and change their answer to the target question since they would have an opportunity to respond to it again on the next page.

The following sections present and discuss the questions given to students in more detail. Note that the screening, target, and each of the analytic support questions were comprised of two parts: a multiple-choice question followed by a short answer question asking students to explain their response. Since the questions were administered on online quizzes graded for completion, many student explanations were so brief that accurate reasoning could not be assessed. Therefore, for the purpose of analyses, student performance on each of the questions was solely based on their multiple-choice selection. However, some student explanations were investigated to gain more insight into student response patterns on the screening and target questions, which will be briefly discussed.

Screening and target questions

The screening and target questions analyzed in this chapter are shown in Fig. 4.4 and come from [29]. The screening question involves a box on a rough floor that has a horizontal force applied to it but the box remains at rest. Students were asked a multiple-choice question

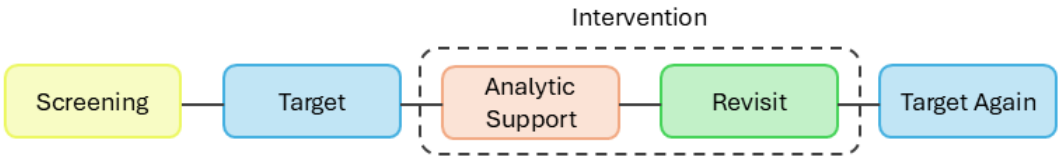


Figure 4.3: Box friction question sequence components.

(question 1) to compare the magnitude of the friction force acting on the box with the applied force. A short answer question followed (question 2) asking them to explain their response.

In order to answer the screening question correctly, students must recognize that since the box is at rest, it must be experiencing balanced forces. Therefore, the force of friction must be equal in magnitude to the applied force.

The target question involves a similar scenario except this time there are two boxes of equal mass experiencing the same horizontal applied force as before but that are resting on different surfaces. The coefficients of static friction between the boxes and each floor are different. The question asked students to compare the magnitudes of the friction forces exerted on each box (question 3), and then explain their response (question 4).

Like the screening question, the target question requires the application of Newton's second law to recognize that each box must have balanced forces. Therefore the force of friction on each box must be equal to 30 N, the magnitude of the horizontal applied force to each box. The correct answer is that the friction forces are equal to each other.

The most common incorrect answer on the target question is that the force of friction is greater for the surface with a greater coefficient of static friction. Sometimes the equation, $F_{fr} \leq \mu_s N$ is used to justify this response, where F_{fr} is the force of static friction, μ_s is the coefficient of static friction, and N is the normal force. While this equation is valid here, the force of static friction is variable depending on the situation. The maximum amount of static friction possible is equal to $\mu_s N$ but could be less. In the case of the target question,

the boxes are at rest and so the force of static friction does not reach its maximum amount for either box, instead matching the applied force to maintain zero acceleration.

Intervention questions

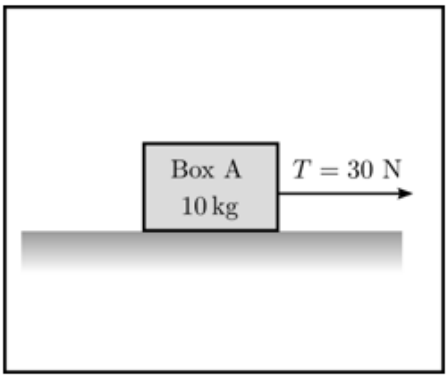
The analytic support questions analyzed in this chapter involved a fictitious student conversation about the target question (see Fig. 4.5). Students were then asked how each of these fictitious students would respond to the target question. Since student 1 is using the equation for static friction to calculate the force of friction on each box, this student would argue that the magnitude of the force of friction on Box A is less than that on Box B. In contrast, student 2 would argue that the magnitude of the friction force is equal for both boxes since friction force must be equivalent to the applied force on each box.

In the analyses that follow, accurate performance on the analytic support questions is defined as correctly answering both of the multiple-choice analytic support questions correctly. If a student responded incorrectly to one or both of the analytic support questions, their overall response to the analytic support questions was considered to be incorrect.

This structure mimics that used in [30], which follows the same three design principles outlined in Sec. 4.4. Regarding design principle (1), this analytic support intends to raise doubt about the importance of the surface when objects remain at rest. This can be seen in student 2's argument. Design principle (2), which focused on cuing relevant content knowledge, is addressed in student 2's reference to Box B remaining at rest, Newton's laws, and balanced forces. Lastly, students are encouraged to engage in process 2 thinking in accordance with design principle (3) since they must consider two different lines of reasoning in order to supply answers to the analytic support questions.

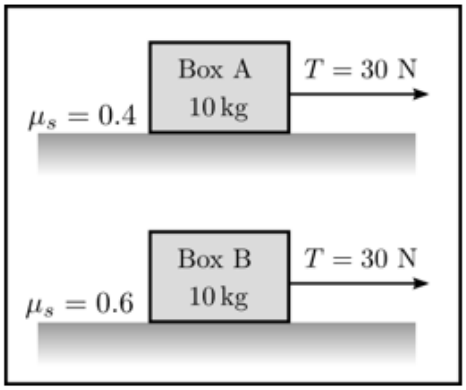
Note that this part of the intervention shares some similarities to prior studies in the box-friction context, but also provides a unique approach. It contains more elements intended to engage process 2 than in [29], including information geared at raising doubt about the common incorrect default model related to the coefficients of friction similar to [12]. In

Box A is initially at rest on a rough floor. A horizontal 30 N force is then applied to the box, as shown below. The box remains at rest.



1. Is the magnitude of the friction force exerted on box A *greater than, less than,* or *equal to* the magnitude of the applied force on box A?
2. Explain.

Suppose the coefficient of static friction between box A and the floor is 0.4, as shown in the diagram below. The coefficient of static friction between box B and a different floor is 0.6, also shown in the diagram below. $m_A = m_B = 10$ kg.



A horizontal 30 N force is applied to each box, and both boxes remain at rest.

3. Is the magnitude of the friction force exerted on box A *greater than, less than,* or *equal to* that exerted on box B?
4. Explain.

Figure 4.4: Box friction screening and target questions.

The students decide to tackle the problem by thinking about box B first. Here are their initial ideas:

Student 1: We know from the equation sheet that static friction F^{s}_{12} can be calculated from $F^{s}_{12} \leq \mu_s F^{n}_{12}$. For box B, μ_s times normal force is 58.8 N, so this must be the value of friction acting on box B.

Student 2: I think 58.8 N is the *maximum* amount of static friction box B can have, but it could be less. If we think about Newton's laws, box B is at rest so it must have balanced forces. That means that the friction force is equal to the applied force.

Figure 4.5: Fictitious student conversation used as part of the box-friction intervention. Directly before the conversation text, students were told that "two students work together to answer question 3 above about the friction forces exerted on the two boxes." The subsequent analytic support questions asked students, "based on the conversation above, how would **student 1** answer question 3" and "how would **student 2** answer question 3?" They were given multiple-choice options and then short answer questions where they could explain their reasoning.

contrast to the reflective intervention implemented in [49], the analytic support questions have students engage in alternative lines of reasoning instead of only reflecting on their own reasoning and reading solutions. The reflexive method used in [49] resulted in larger student improvements on the target question than their reflective approach and the authors suggest that reflexive interventions would be better suited in contexts with salient distracting features. However, a reflective intervention method using alternative lines of reasoning similar to the one presented here has been shown to be effective in a different context involving salient distracting features [47]. The intervention presented in this dissertation tests the type of reflective approach involving alternative lines of reasoning on the well-researched box-friction questions in an effort to establish a proof of concept for the validation methods developed in [41] and [30].

In addition to the analytic support questions discussed above, most sequences included a "revisit" question adapted from [46] as shown in Fig. 4.6. This question was intended to address design priciple (1) by raising doubt in students' response to the target question and nudge them towards revisiting their thinking on the target question. It acted as a measure of students' desire to revisit their thinking. Students who indicated that they would answer the target question differently than before (answer choice a) or that they would add to or change their explanation (answer choice b) were considered to have a desire to revisit.

4.7 Results and Discussion

In the following sections, a discussion of the data is included and analyses are provided to validate the screening-target question pair as well as the intervention questions. Student response patterns are investigated as well as various analyses involving students' cognitive reflection scores. Note that results related to students' three-item CRT scores are presented. While analyses were also conducted using seven-item CRT scores, the same trends were generally observed, and the three-item CRT seemed sufficient for discriminating students' cognitive reflection skills. Any future mention of the CRT is specifically in reference to

Answering questions 5 through 8 may (or may not) have led you to think more about question 3 or even reconsider your answer to question 3. Below, reflect on how (if at all) your thinking has changed.

a. After answering questions 5 through 8, I would now answer question 3 differently than before.
b. After answering questions 5 through 8, I would answer question 3 the same as before, but I would add to or change my explanation somewhat.
c. After answering questions 5 through 8, my answer and explanation on question 3 are still the same.

Figure 4.6: Revisit question presented after the analytic support questions (5 through 8). Note that question 5 (multiple choice) asked how a fictitious student 1 would answer the target question followed by an opportunity to explain reasoning in question 6. Questions 7 and 8 were the same as 5 and 6 except with regards to fictitious student 2. Note that question 3 was repeated within the analytic support questions, so it was fresh in students' minds when answering the revisit question.

the three-item version. Following detailed results and discussion regarding the box-friction context, a comparison to results from the pulse sequence is described.

4.7.1 Discussion of the data

As part of the iterative process of intervention development, various forms of the question sequence were administered during five different quarters from Autumn of 2022 to Winter of 2025. The question sequences (six in all) were given at different times within the instructional context each quarter. Fig. 4.7 shows how each sequence fell with respect to relevant lectures and tutorials [51]. Note that in the Spring of 2024, there were two instances during the quarter where students were presented with a question sequence.

Fig. 4.8 shows the components of each iteration of the question sequence. During the first quarter, which I will denote as Qtr 1, only the original screening and target questions as seen in Fig. 4.4 were given to students to establish the validity of the screening-target pair with a sample population before developing an intervention in this context. In Fig. 4.8, the

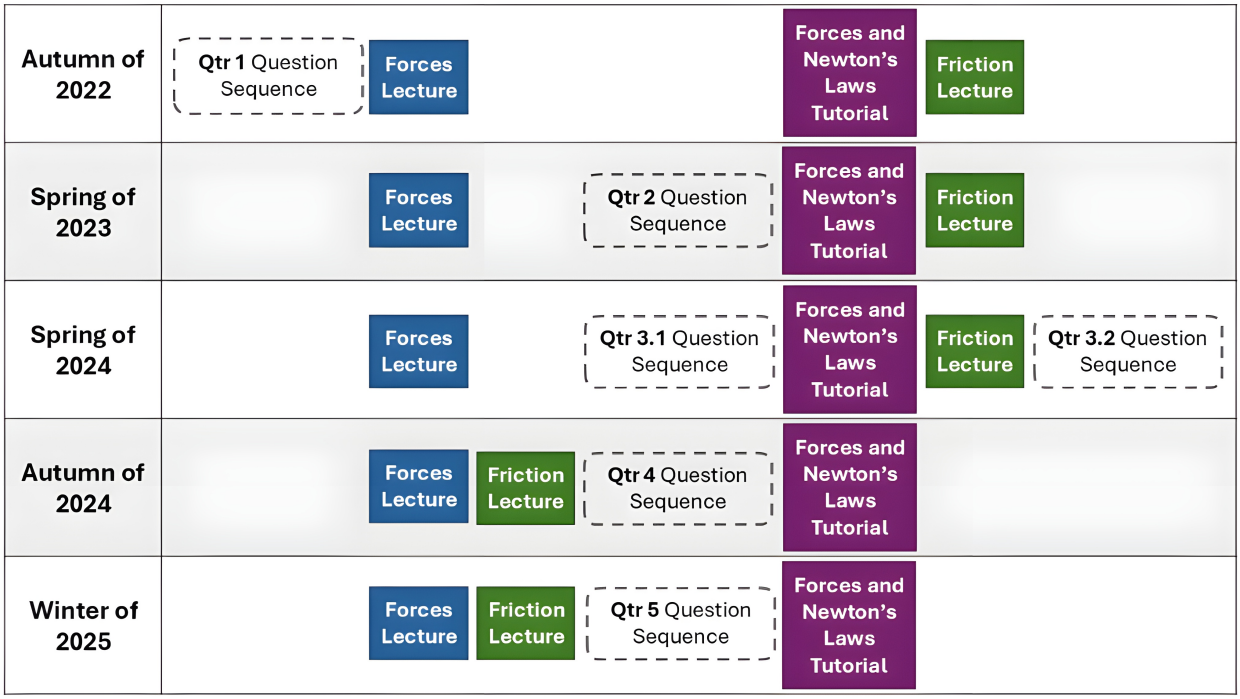


Figure 4.7: A variation of the question sequence was given during five academic quarters. The diagram shows the instructional timing for each administration.

original target question is referred to as the "2-box target" question to distinguish it from a "3-box target" version discussed below that was administered on other question sequences. (The three-box target question can be viewed in Appendix B.)

In Qtr 2, two versions of the question sequence were given on the same online quiz where students were assigned one of the two versions randomly. Both versions had most questions in common except for the target question. One group of students received the original two-box target question while the other group received a three-box target question. The three-box question consisted of the same setup as the original version with an additional third box on a surface with a coefficient of friction equal to 0.6 and an applied force of 20 N. The question asked students to rank the magnitudes of the friction forces on each box. This additional version of the target question was given because it probed student thinking in another capacity, namely about how the frictional force would compare for identical boxes

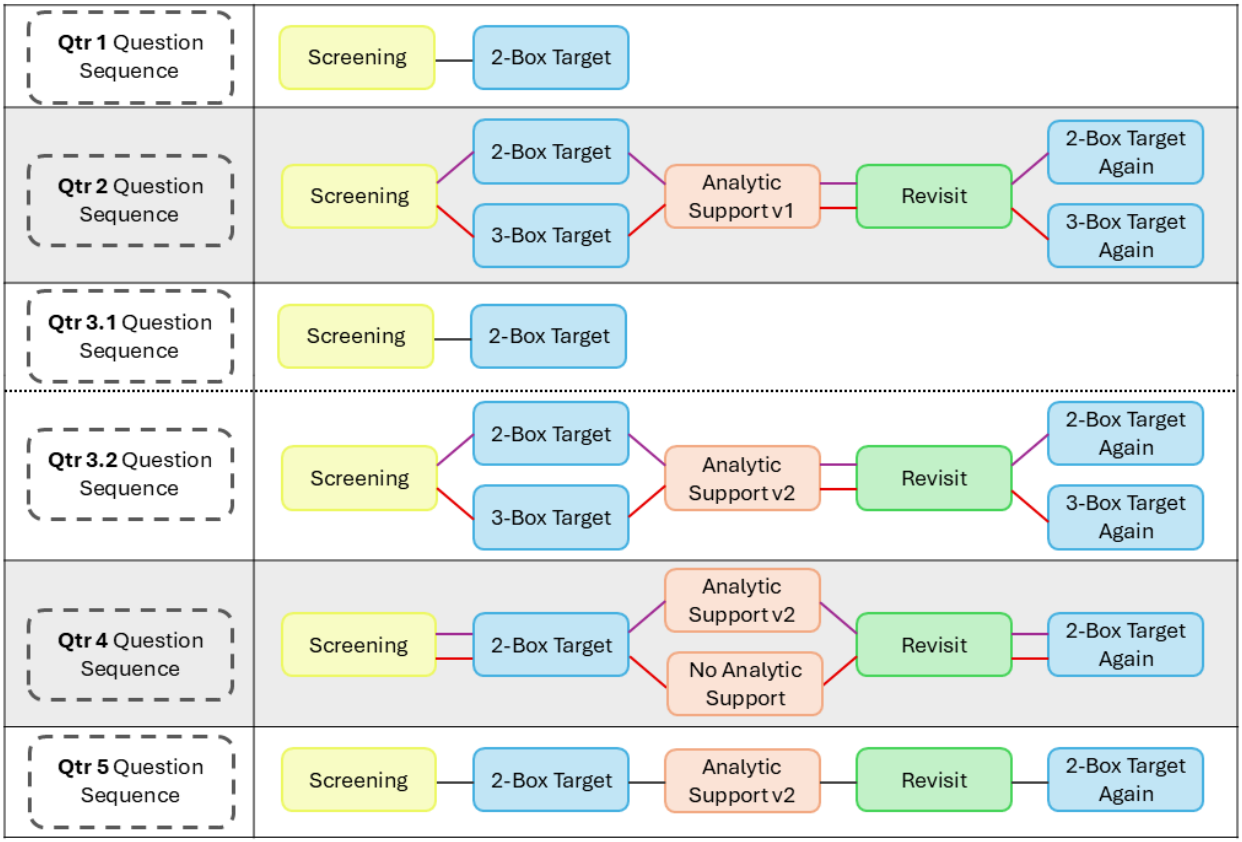


Figure 4.8: The components of each administration of the question sequence. Some iterations included two versions. The screening and two-box target questions can be found in Fig. 4.4, the analytic support v2 questions can be found in Fig. 4.5, and the revisit question is displayed in Fig. 4.6. The three-box target question and analytic support v1 questions are located in Appendices B and C respectively.

at rest with differing magnitudes of applied force.

The first iteration of the analytic support questions (denoted as v1 in Fig. 4.8) was implemented on the Qtr 2 sequence. This iteration involved a fictitious student dialogue where the conversation focused on the variable nature of the static friction force and the amount of applied force needed for the box to move. (See Appendix C for the analytic support v1 student dialogue.) Students were asked how each fictional student would respond to the target question (two-box or three-box depending on the question sequence version).

In Spring of 2024, two question sequences were given at separate times during the academic quarter. The Qtr 3.1 sequence only contained the original screening and target questions, which were already contained on an online quiz as part of the course. The Qtr 3.2 question sequence administered later in the quarter, which included an intervention, was a research-informed choice motivated by the desire to assess students after they had received all in-class instruction related to relevant topics. The Qtr 3.2 sequence ran as two versions identical to those used in Qtr 2 except for modifications to the analytic support questions. This iteration of the analytic support questions (v2) involved a fictitious student conversation (see Fig. 4.5) that focused more on ideas related to Newton's laws, which was intended to better steer students away from focusing on the surfaces and more directly cue ideas about balanced forces and Newton's laws. This choice was made with the aim to ultimately improve student performance on the target question to a greater degree than the previous iteration.

On the Qtr 4 question sequence, a decision was made about which target question to continue to administer for subsequent research. In this case, only the two-box target question was given. While student response patterns to both the two-box and three-box versions of the target question supported the two assumptions for screening-target pairs (outlined above), student performance was worse on the three-box version. Additionally, analyses of student response patterns to the original screening-target pair is documented in the literature, making analyses conducted with the two-box version in this research more comparable.

While only the two-box target question was used, two versions of the question sequence were administered in Qtr 4, this time with and without the analytic support v2 questions. The intent was to see if merely asking students if they wanted to revisit their response to the target question was enough to result in significant student improvement on the target question, particularly in comparison to the sequence version also containing the analytic support questions.

Finally, after observing that the analytic support questions in combination with the opportunity to revisit was indeed resulting in greater student improvement on the target question compared to excluding the analytic support questions in Qtr 4, a single version of the question sequence was given in Qtr 5. This sequence was identical to the Qtr 4 version containing the analytic support questions.

Since all question sequences included the original screening and two-box target questions, student response data for these questions on each sequence is presented in table 4.1 for comparison. Three measures are presented: overall student performance on the screening question, overall student performance on the target question, and performance on the target question for only those students who answered the screening question correctly. The last measure was included because one assumption is that student accuracy on the target question is dependent on their application of relevant content knowledge on the screening question. (Note that the number of student respondents listed in the table corresponds to only those students who received a question sequence containing the two-box version of the target question.)

It was apparent that performance appeared to generally improve on each subsequent administration of the sequence, which was investigated in further detail. Most of the sequences were given around the same time within the instructional context with the exception of Qtr 3.2, which was given after students had completed the research-based *Tutorials in Introductory Physics: Forces and Newton's Laws* [51]. The University of Washington *Tutorials in Introductory Physics* have been shown in many studies to improve students' conceptual mastery

Table 4.1: Screening and two-box target question performance for each individual administration of the questions. (Each N represents only the number of students who responded to the two-box version of the target question.

Question sequence	N	% Screening correct	% Target correct	% of Screening correct with target correct
Qtr 1	260	56	47	63
Qtr 2	62	68	48	67
Qtr 3.1	454	71	56	68
Qtr 3.2	232	78	64	72
Qtr 4	332	76	74	86
Qtr 5	384	80	79	89

of various physics topics [52, 53, 54, 55, 56, 57, 58, 59], and so potential improvement on the Qtr 3.2 administration is expected. It is also possible that students may have improved to some extent just by nature of seeing the same two questions again.

Chi-square test comparisons were run across all other quarters to determine if there were statistically significant differences in screening and target performance between any two academic quarters. A Bonferroni correction was used due to multiple comparisons such that the α level was set at 0.005 (0.05 divided by the number of tests performed). The first significant result showed a small-effect size improvement on the screening question in Qtr 3.1 compared to Qtr 1 [$p = 6.42 \times 10^{-5}$, Cramér's V effect size = 0.15 (small)]. This could be due to instructional timing given that students in Qtr 1 had not been exposed to a lecture on forces before answering the questions. There was no difference in performance on the target question between these two quarters, which is not surprising since the screening question was designed to be a straightforward test of content-knowledge while the target question was intended to not only require content-knowledge but cognitive reflection as well. One

lecture on forces would not be expected to improve students' cognitive reflection skills.

Every other statistically significant comparison showed that performance on each of the questions in Qtr 4 and Qtr 5 was better than that in various preceding quarters (small Cramér's V effect sizes in all cases except a medium effect size improvement on Qtr 5 target performance compared to Qtr 1). (The only nonsignificant comparisons were between Qtr 2 screening performance and both Qtrs 4 and 5, and between Qtr 3.1 screening performance and Qtr 4.) More generally, student performance has been observed to significantly improve on various online quiz assignments during Qtrs 4 and 5, which is currently under investigation. This sudden jump in response accuracy during Qtrs 4 and 5 has been proposed to stem from the increased accessibility and popularity of artificial intelligence (AI) chatbots, such as ChatGPT, which students have been known to use for providing answers to questions such as these. Note that GPT-4o, which became publicly available in May of 2024, answers both the screening and target questions correctly.

Given the popularity of AI chatbots prior to Autumn of 2024, there is a chance that some students were using AI chatbots in Qtr 3.1 as well, but it doesn't appear to have largely affected the data, especially since there was no improvement on the target question compared to previous quarters.

Since data from Qtrs 4 and 5 are potentially not representative of student thinking, the analyses discussed here focus on the previous quarters. The validation of the screening and target questions utilizes combined data from Qtrs 1, 2, and 3.1 since these administrations shared similar instructional timing and similar target question performance for content-proficient students (63%, 67%, and 68% correct respectively). I refer to this set of data simply as the "combined data" moving forward. An analysis of the screening and target questions is also provided for the Qtr 3.2 sequence separately. The intervention analysis is conducted using the Qtr 3.2 data since this was the only quarter where the intervention was administered prior to the last two quarters.

4.7.2 Validation of the screening and target questions

Below, the screening and target questions are examined for (1) their ability to identify students with relevant conceptual understanding and (2) their consistency with the DPTs interpretation that, for content-proficient students, responding to the target question is largely an issue of engaging in cognitive reflection on an incorrect provisional model. Recall that content-proficient students are considered to be those who have responded correctly to the screening question.

First, these two assumptions are tested on the combined data from Qtrs 1, 2, and 3.1. Next, these assumptions are evaluated for the Qtr 3.2 administration of the question sequence individually. The Qtr 3.2 data is analyzed individually because it involves students' second responses to the screening and target questions (after seeing them on the Qtr 3.1 question sequence). The Qtr 3.2 intervention data is used for intervention analysis as well, and so it was of interest whether the screening and target questions were functioning as intended for this particular sequence. Lastly, a discussion of all the screening-target analyses are discussed.

Testing assumption (1)

In section 4.4, two main DPTs-based assumptions are described. In this section, analyses used to evaluate assumption (1) are discussed. Recall that this assumption holds that "the screening question accurately identifies students as having, or not having, the conceptual understanding needed to correctly answer the target question." If this assumption is true, correctly responding to the target question should depend on correctly responding to the screening question. However, correctly responding to the screening question should not necessarily imply correctly answering the target question. This is because the target question presents coefficients of friction which act as a salient distracting feature. For students who first think of an incorrect idea, potentially based on the coefficients of friction, arriving at a correct response depends not only on their conceptual understanding but their engagement

Table 4.2: Contingency table of student performance on screening and target questions, which were administered on a quiz covering forces and Newton's laws during Qtrs 1, 2, and 3.1.

	Screening incorrect	Screening correct	Total
Target incorrect	200	172	372
Target correct	66	337	403
Total	266	509	775

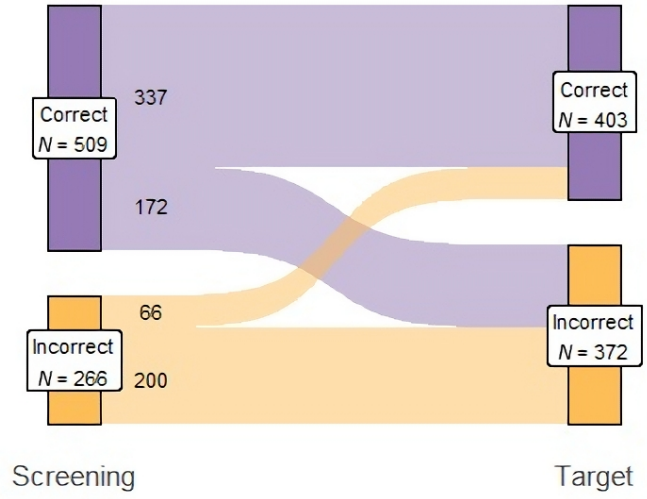


Figure 4.9: Sankey diagram showing student performance on the screening and target questions for combined data.

of cognitive reflection as well.

To test the relationship between the screening and target questions, a chi-square test was conducted comparing the screening question performance for the group of students who answered the target question correctly and those who answered the target question incorrectly (see Table 4.2). The data is also presented in a Sankey diagram in Fig. 4.9. The data indicate that those who answered the target question correctly were more likely to answer the screening question correctly [$p = 6.51 \times 10^{-28}$, Cramér's V effect size $= 0.39$ (medium)], which supports the predicted relationship. Using the same data from Table 4.2, a risk ratio

was calculated to determine how many times more likely this was the case.

$$\begin{aligned}\text{Risk Ratio} &= \frac{\text{(Screening correct \& target correct)/Target correct}}{\text{(Screening correct \& target incorrect)/Target incorrect}} \\ &= \frac{337/403}{172/372} = \frac{0.84}{0.46} = 1.81\end{aligned} \tag{4.1}$$

Students who were correct on the target question were 1.81 times more likely to perform successfully on the screening question than those who failed to answer the target question correctly.

In addition to the above tests, the data was analyzed using the methods described in [35] to more carefully evaluate the existence of the expected hierarchy, in this case that correctly answering the target question necessarily implies correctly answering the screening question. This is supported by the fact that 84% of students who performed successfully on the target question also answered the screening question correctly. Equivalently, within this hierarchy one would expect that incorrectly answering the screening question implies incorrectly answering the target question. Here, 75% of students who failed to answer the screening question correctly also failed to correctly respond to the target question, which supports this relationship. It is not necessary for those who answer the screening question correctly to go on to answer the target question correctly, and this is observed by the 172 students who exhibited this answering pattern. For a perfect hierarchy, one would expect no students to be able to answer the target question correctly without also demonstrating relevant content knowledge by correctly answering the screening question. However, there are 66 students who fall into this category.

This could be partially due to random guessing. There is also a possibility that students may have chosen the correct answer to the target question for incorrect reasons. The most common incorrect answer to the screening question is that in order for the box to remain at

rest, the friction force must be greater than the applied force of 30 N. For example, on the screening question one student stated, “the box is not moving, so there is more force from the friction than the force applied to the box.” A portion of students with this notion apply it to the target question as follows: both boxes must have frictional forces that are equal to each other (but greater than 30 N) because they are resisting the same applied force in order to stay at rest. The same student quoted above responded to the target question by saying, “The force is similar because it only requires the same amount of force to stop the box from moving, meaning they will have the same force.” In this way, some students chose a correct answer for the target question but did not demonstrate appropriate understanding of balanced forces on the screening question.

This only has minor implications on the subsequent analyses. Some analyses conducted in the sections below compare students based on target performance to determine if there are any differences in these groups. There may be some students who answer the target question correctly with incorrect reasoning who should be categorized with students who failed to answer the target question correctly but there do not appear to be enough instances of this to make a difference in the results.

Overall, the data support assumption (1), and the screening question is generally identifying students who have the content knowledge necessary to successfully answer the target question. It is worth noting that the screening-target pair is not a perfect ”sorter” of those with conceptual understanding. Of course there may be some students who answer the screening question correctly who do not have a complete or robust set of mindware necessary for correctly answering the target question since students do not necessarily need to think about static friction while answering the screening question but may need to employ a greater understanding of this concept on the target question. Even still, the observed hierarchy aligns with the general trends expected from a dual-process theories perspective.

Table 4.3: Contingency table of content-proficient student performance on the target question based on high or low CRT score for combined data.

	Low CRT (0-1)	High CRT (2-3)	Total
Target correct	46	291	337
Target incorrect	39	133	172
Total	85	424	509

Testing assumption (2)

In validating the screening and target questions, assumption (2) was also tested, that for content-proficient students, responding to the target question largely involves mediating an incorrect intuitive thought with cognitive reflection. If true, one would expect to see (a) an association between content-proficient students' CRT scores and their performance on the target question. One would also anticipate that (b) CRT score is predictive of performance on the target question for content-proficient students.

Content-proficient students' CRT scores are associated with target performance

One would expect that for students who demonstrated adequate content knowledge on the screening question, those with high CRT scores would outperform those with low CRT scores on the target question. A score of 0 or 1 was considered low while a score of 2 or 3 was considered high on the CRT.

The combined screening and target question data supported this hypothesis. Table 4.3 shows a contingency table comparing the number of content-proficient students with low and high CRT scores based on their performance on the target question. A chi-square test indicates there is a small effect size association between success on the target question and high CRT score for content-proficient students [$p = 0.00982$, Cramér's V effect size $= 0.11$ (small)].

To determine the relative likelihood of correctly answering the target question for content-

Table 4.4: Contingency table showing difference in low and high CRT scores between content-proficient and noncontent-proficient students for combined data.

	Low CRT (0-1)	High CRT (2-3)	Total
Screening incorrect	75	191	266
Screening correct	85	424	509
Total	160	615	775

proficient students with differing CRT score levels, the risk ratio shown below was calculated.

$$\begin{aligned}\text{Risk Ratio} &= \frac{\text{(Target correct \& content proficient high CRT)/Content proficient high CRT}}{\text{(Target correct \& content proficient low CRT)/Content proficient low CRT}} \\ &= \frac{291/424}{46/85} = \frac{0.69}{0.54} = 1.27\end{aligned} \tag{4.2}$$

This shows that content-proficient students with high CRT scores were 1.27 times more likely to respond successfully to the target question compared to content-proficient students with low CRT scores.

This result supports assumption (2). It appears that, for students who possess the relevant mindware, a tendency to mediate intuitive ideas with analytic thinking is advantageous for correctly answering the target question. Therefore, not only does the target question seem to depend on content knowledge assessed via the screening question, but cognitive reflection as well, which is consistent with a dual-process theories interpretation.

However, the results only weakly support assumption (2) given the small effect size. This may be attributed to the fact that performance on the screening question was somewhat dependent on cognitive reflection as well. For all students, there was a small effect size association that students with high CRT scores were more likely to answer the screening question correctly than those with low CRT scores [$p = 0.000174$, Cramér's V effect size =

0.13 (small)]. Table 4.4 provides the combined data used for this analysis. Since students who correctly answered the screening question generally had higher CRT scores, when comparing their performance on the target question to their CRT scores, a smaller effect size difference could be observed due to the limited variability in CRT scores.

Content-proficient students' CRT scores predict target performance Assumption (2) can also be investigated by performing a binary logistic regression to see if CRT score was a predictor of answering the target question correctly for content-proficient students. If this assumption is correct, it follows that an increase in CRT score for content-proficient students would increase the likelihood of success on the target question.

Fig. 4.10 shows a binary logistic regression generated from the combined screening and target data. Each point (jittered for visualization purposes) represents the CRT score and target performance of a single content-proficient student, where 0 on the y-axis represents an incorrect answer to the target question while 1 on the y-axis represents a correct answer. The regression represents the probability of responding correctly to the target question as a function of CRT score.

The regression shows that, for every one-point increase in CRT score, the odds of answering the target question correctly increased by 1.59 for content-proficient students (95% CI [1.30, 1.96]). This corroborates the previous analysis in supporting assumption (2).

Validation of Qtr 3.2 screening and target questions

While the combined screening and target data (consisting of student responses on the Qtr 1, Qtr 2, and Qtr 3.1 sequences) allowed for validation of the screening and target questions in the box-friction context, the same assumptions were also tested individually on the Qtr 3.2 administration of the sequence. Data from Qtr 3.2 also included student responses to an entire intervention sequence as well, results of which will be discussed below in Sec. 4.7.3.

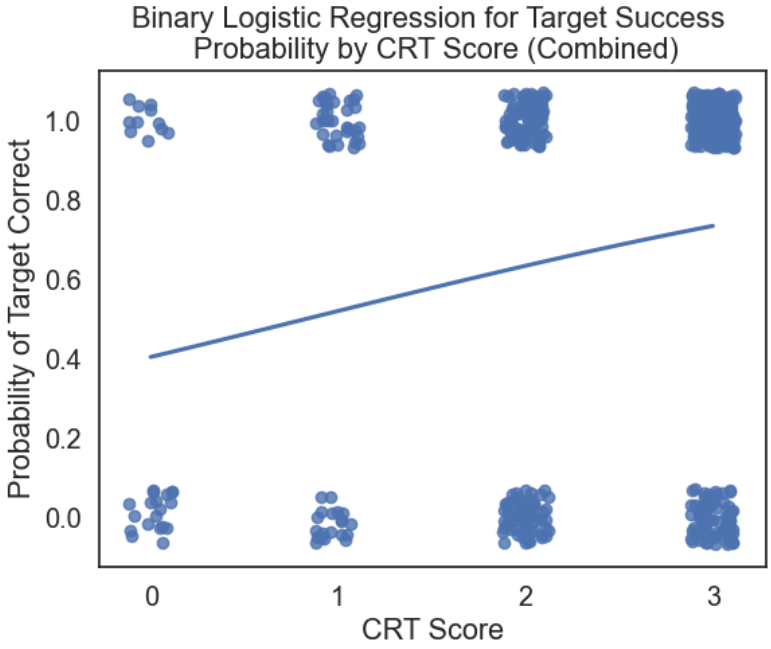


Figure 4.10: Binary logistic regression showing the probability of content-proficient students responding correctly to the target question depending on their CRT score for combined data ($p = 9.0 \times 10^{-6}$, odds ratio = 1.59, 95% CI [1.30, 1.96]). Each data point represents one content-proficient students' CRT score (exactly 0, 1, 2, or 3) and their performance on the target question (0 = incorrect, 1 = correct). The data points have been jittered for visualization purposes.

Table 4.5: Contingency table of student performance on screening and target questions given as part of the complete question sequence administered during Qtr 3.2.

	Screening incorrect	Screening correct	Total
Target incorrect	34	50	84
Target correct	18	130	148
Total	52	180	232

Therefore, it was relevant to evaluate whether the screening and target question were functioning as intended on this individual administration as well.

Testing assumption (1) for Qtr 3.2 data Similar to the analysis reported in section 4.7.2, students' screening and target performance was evaluated to gain insight into whether the screening question was correctly identifying students with mindware necessary to correctly answer the target question. A Sankey diagram of screening and target performance for the late administration of the sequence is displayed in Fig. 4.11, and a contingency table of the data is presented in Table 4.5. A chi-square test shows that students who answered the target question correctly were more likely to have also succeeded on the screening question [$p = 6.69 \times 10^{-7}$, Cramér's V effect size = 0.33 (medium)]. In fact, the following risk ratio shows that students who correctly responded to the target question were 1.48 times more likely to answer the screening question correctly as compared to students who incorrectly responded to the target question.

$$\begin{aligned} \text{Risk Ratio} &= \frac{\text{(Screening correct \& target correct)/Target correct}}{\text{(Screening correct \& target incorrect)/Target incorrect}} \\ &= \frac{130/148}{50/84} = \frac{0.88}{0.60} = 1.48 \end{aligned} \tag{4.3}$$

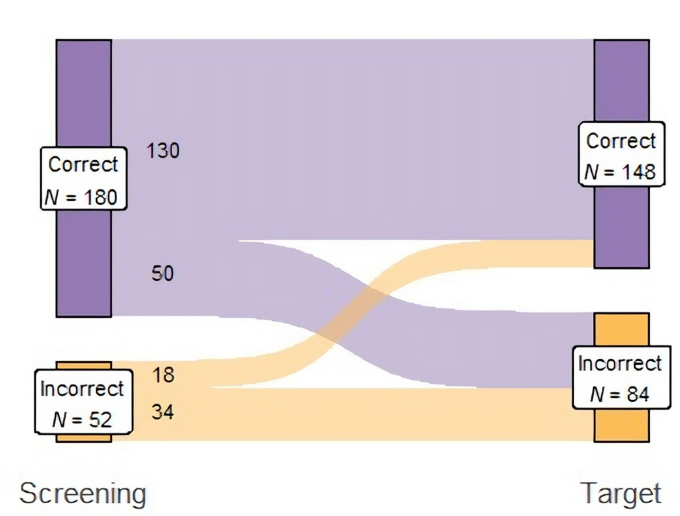


Figure 4.11: Sankey diagram showing student performance on the screening and target questions on the complete question sequence during Qtr 3.2.

These results support the assumption that the screening question is identifying students who have the relevant content knowledge needed to succeed on the target question.

Using the methods from [35] to assess the presence of a hierarchy, the statement that correctly answering the target question necessarily implies correctly answering the screening question is supported by the fact that 88% of students who correctly answered the target question also correctly answered the screening question. In addition, 65% of students who incorrectly answered the screening question went on to incorrectly answer the target question, supporting the idea that incorrectly answering the screening question implies responding incorrectly to the target question as well. Those who succeed on the screening question do not necessarily have to correctly respond to the target question under this hierarchy, and there were 50 students who fell into this category. Lastly, a perfect hierarchy would show no students succeeding on the target question who failed to answer the screening question correctly. There were 18 students who exhibited this answering pattern in the data, but as discussed earlier this could be due in part to random guessing and the possibility of answering the target question correctly with incorrect reasoning that would be associated with answering

Table 4.6: Contingency table showing difference in low and high CRT scores between content-proficient students who answered the target question differently during Qtr 3.2.

	Low CRT (0-1)	High CRT (2-3)	Total
Target correct	18	112	130
Target incorrect	13	37	50
Total	31	149	180

the screening question incorrectly (see Sec. 4.7.2 for a discussion of this).

Testing assumption (2) for Qtr 3.2 data In addition to testing the screening and target pair for their ability to identify students demonstrating mindware in this particular data set, assumption (2) was also tested, that success on the target question is largely dependent on engaging in process 2 thinking to override an incorrect intuition. The same analyses as in Sec. 4.7.2 was used. Under this assumption, one would expect content-proficient students with a greater propensity for cognitive reflection (as seen from high CRT scores) to perform better on the target question than those with low CRT scores.

A chi-square test was conducted to compare content-proficient students' CRT scores and target performance (see data in Table 4.6). However, the results showed no significant difference between the target performance of content-proficient students based on CRT score ($p = 0.053$). As before, I checked to see if there was a dependency on cognitive reflection for success on the screening question that could be diluting the above results. A chi-square test to this effect using the data shown in Table 4.7 did not show any significant results ($p = 0.52$).

The fact that the Qtr 3.2 administration of the screening and target questions did not show significant results is consistent with other individual administrations including Qtrs 1, 2, and 3.1. However, when combined, the Qtr 1, 2, and 3.1 results still contributed to significant

Table 4.7: Contingency table showing difference in low and high CRT scores between content-proficient and noncontent-proficient students on the tutorial pretest given during Qtr 3.2.

	Low CRT (0-1)	High CRT (2-3)	Total
Screening incorrect	11	41	52
Screening correct	31	149	180
Total	42	190	232

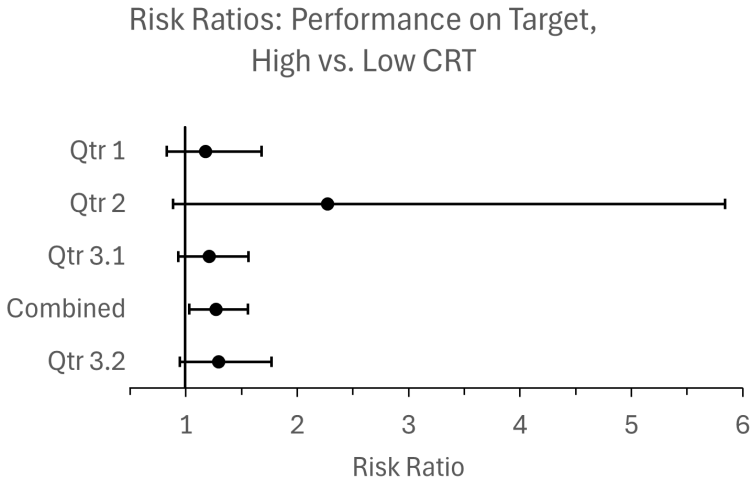


Figure 4.12: Forest plot of risk ratios with 95% confidence intervals for individual and combined data. The risk ratios indicate how many times more likely content-proficient students with high CRT scores were to answer the target question correctly compared to content-proficient students with low CRT scores.

results. Fig. 4.12 shows a forest plot of risk ratios with 95% confidence intervals (determined using the Katz method) for the likelihood of content-proficient students correctly answering the target question with high CRT scores compared to those with low CRT scores. Even though not all individual results are significant, there is a general trend indicating that content-proficient students with high CRT scores are more likely to answer the target question correctly than those with low CRT scores. The Qtr 3.2 data falls in line with this trend.

Additionally, the forest plot shown in Fig. 4.13 gives risk ratios with 95% confidence intervals

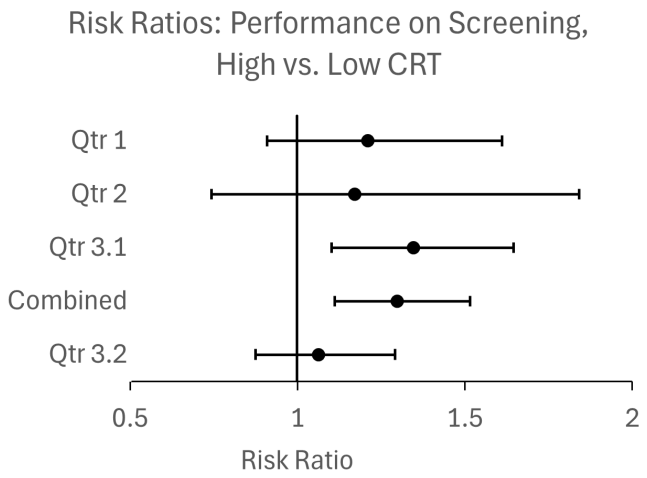


Figure 4.13: Forest plot of risk ratios with 95% confidence intervals for each quiz administration of the screening and target questions in addition to combined data. The ratios represent how many times more likely it was for students with high CRT scores to correctly answer the screening question compared to students with low CRT scores.

representing how much more likely it was for students with high CRT scores to answer the screening question correctly compared to students with low CRT scores.

The results for Qtr 3.2 are consistent with the general trend among all the data that there may be some dependence on cognitive reflection for success on the screening question, which could be contributing to the appearance of no CRT-dependence for content-proficient students' success on the target question. This may not be a large enough data set to be conclusive.

However, a binary logistic regression intended to determine the predictive power of content-proficient students' CRT scores on their target question performance gives more insight into the validity of assumption (2) (see Fig. 4.14). As expected, the higher a content-proficient student's CRT score, the greater their probability of success on the target question. Specifically, for every one point increase in CRT score, the odds of a content-proficient student correctly responding to the target question increased by 1.52 (95% CI [1.06, 2.20]). This provides credence to the assumption that performance on the target question is largely

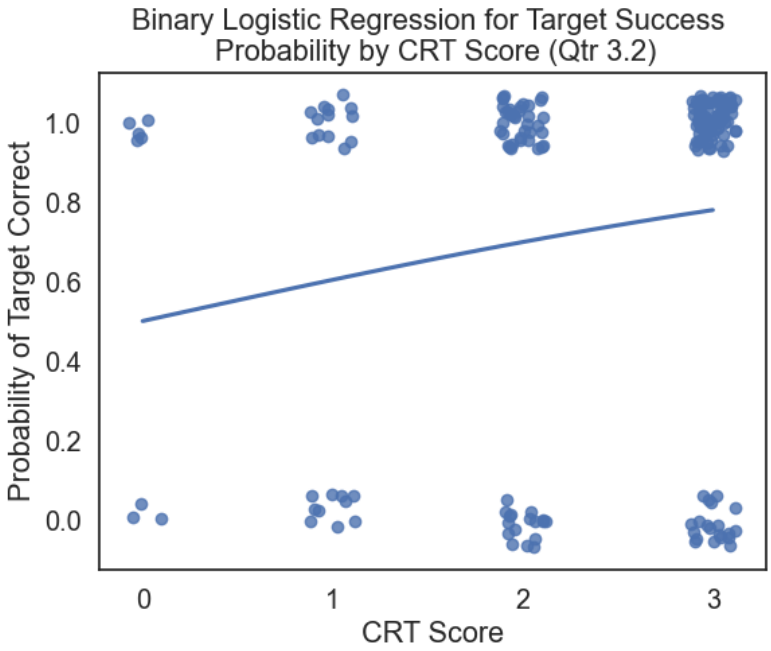


Figure 4.14: Binary logistic regression for content-proficient students who completed the Qtr 3.2 question sequence. The probability of responding correctly to the target question depended on students' CRT scores ($p = 0.024$, odds ratio = 1.52, 95% CI [1.06, 2.20]). Each data point (jittered for visualization purposes) represents one content-proficient students' CRT score and their performance on the target question (0 = incorrect, 1 = correct).

a matter of reflecting on an incorrect process 1 model.

While the chi-square tests used to evaluate assumption (2) were inconclusive, the general trend is consistent with previous data in support of assumption (2). This, in combination with a binary logistic regression showing CRT score as predictive of target performance for content-proficient students, corroborates the validity of assumption (2) for Qtr 3.2.

Discussion of screening-target validation

Overall, results suggest that the screening and target questions are behaving more or less as intended for combined data and Qtr 3.2 data. It appears that dual-process theories can explain student response patterns on the box-friction screening-target pair. Namely, the screening question is acting as an indication of the presence of certain content knowledge [assumption (1)], while the target question appears to depend on the same content knowledge in addition to cognitive reflection on incorrect process 1 intuition [assumption (2)].

It is worth discussing that assumption (2) is only weakly supported by the data. There are many confounding factors that could be contributing to this.

(1) Some students (even with high CRT scores) may have felt so strongly about an incorrect provisional model generated by process 1 that they did not recognize the need to reflect, and therefore did not engage in process 2 thinking. In fact, it has been suggested in other another study of the same question pair that those who are drawn to the salient distracting feature of the coefficients of friction likely have a strong feeling of rightness about the common incorrect intuitive model [12].

(2) Even if students did engage in process 2 thinking (presumably a group of students with higher CRT scores), they may have (incorrectly) justified the notion that "a greater coefficient of friction means greater friction" using the equation $F_{fr} \leq \mu_s N$. This equation is often cited by students to confirm the model supported by the notion above, and some students have been known to dismiss qualitative reasoning if it appears

to disagree with a mathematical formula, placing formulas in a position of higher importance during reasoning [22].

(3) There is a chance that some students' process 1 generated correct ideas when presented with the target question, and so their tendency to reflect was not relevant to their performance on this question.

(4) There is also a chance that some students used the assistance of AI chatbots to complete the question sequence. If used on the target question, success on the target question would be independent of these students' mindware or cognitive reflection skills.

(5) The cognitive reflection test may also not be a distinguishing measure of cognitive reflection as it correlates with other measures of academic prowess [21, 36, 37, 38].

(6) The cognitive reflection test may require a different kind of cognitive reflection than that needed on the box-friction target question.

(7) Students may have found the presence of the CRT questions on a course quiz to be incongruous with other content, leading them to engage in higher reflection than they would on physics questions in the same setting. In this case, some students with high CRT scores may not have engaged in the same level of cognitive reflection on the target question, leading to a weak association between CRT score and target question performance.

I suspect that factors (1) and (2) are largely affecting the data here, especially given that other research indicates student persistence of incorrect intuitive ideas generated by process 1 for this particular screening-target pair [29, 22].

4.7.3 Evaluation of the intervention

Recall the three design principles considered for producing a DPTs-based intervention centered on supporting process 2 thinking, which is repeated below (see Sec. 4.4):

Principle (1) For students experiencing reasoning hazard A on the target question, they must experience doubt in their process 1 model in order to recognize the need to reflect.

Principle (2) Relevant content knowledge and skills needed to successfully navigate the path of sustained effort must be easily accessible.

Principle (3) Students must have sufficient motivation to engage in productive analytic thinking in order to produce a correct model.

Analyses of the ability of the intervention (as it appeared on the Qtr 3.2 question sequence) to embody the first and third principles is discussed in the following sections. The second principle is not easily measurable since activated pieces of knowledge are internal. It is not guaranteed that students would include cued knowledge in their explanations to the analytic support questions or second attempt at the target question, especially since the questions were given on an online quiz graded for completion and so student explanations were often not robust or complete.

First, design principle (1) is investigated. Second, design principle (3) is evaluated. Third, analyses investigating how performance on the analytic support questions is related to design principles (1) and (3) as well as whether the intervention is equally benefiting students regardless of their cognitive reflection skills is included. Lastly, a summary of the intervention results is discussed.

Investigation of design principle (1)

If the intervention successfully caused students to doubt their initial incorrect ideas on the target question, one would expect that a large percentage of students who incorrectly answered the target question would consider changing their answer and/or explanation to the target question. This was measured using their responses to the revisit question.

Of the 84 students who answered the target question incorrectly on the Qtr 3.2 question

sequence, 64 (76%) indicated a desire to revisit their thinking on the target question by selecting either that they would now answer the target question differently than before or that they would add to or change their explanation to the target question. Given that this is a majority of students in this category, it is plausible that the intervention caused students to consider reflecting on the target question.

In Qtr 4, two versions of the sequence were implemented: one that contained the complete intervention sequence and one that excluded the fictitious student conversation and associated questions in order to investigate the extent to which the analytic support questions were responsible for various student response patterns. While this data is likely compromised by student AI chatbot usage, students who responded incorrectly on the target question presumably did not use an AI chatbot to assist them with their response. Therefore, a comparison of students' desire to revisit the target question is included based on the version of the question sequence they encountered.

Without the analytic support questions, only 48% of students who answered the target question incorrectly indicated a desire to revisit their thinking on the target question. However, 72% of students who completed the analytic support questions after answering the target question incorrectly were willing the revisit their thinking. A contingency table showing this data can be found in Table 4.8. A chi-square test was conducted to determine if there was a significant difference between these two groups and the results indicate a small-effect-size difference that students given the analytic support questions were more likely to select an option on the revisit question consistent with a desire to revisit their answer and/or explanation to the target question [$p = 0.025$, Cramér's V effect size $= 0.24$ (small)].

In fact, the following risk ratio shows that students who saw the analytic support questions were 1.50 times more likely to want to revisit their reasoning on the target question than those who did not receive the analytic support questions. Note that this only includes the group of students who answered the original instance of the target question incorrectly.

Table 4.8: Contingency table showing data for students who answered the target question incorrectly during Qtr 4. The table compares students' desire to revisit their thinking on the target question (obtained from their response to the revisit question) based on the version of the question sequence they completed, either with or without the analytic support questions questions associated with the fictitious student conversation.

	Wanted to revisit	Did not want to revisit	Total
No analytic support	24	26	50
Analytic support	26	10	36
Total	50	36	86

$$\begin{aligned}\text{Risk Ratio} &= \frac{\text{(Wanted to revisit \& analytic support)/Analytic support}}{\text{(Wanted to revisit \& no analytic support)/No analytic support}} \\ &= \frac{26/36}{24/50} = \frac{0.72}{0.48} = 1.50\end{aligned} \tag{4.4}$$

This difference suggests that, for students who failed to respond correctly to the target question, the analytic support questions prompted recognition of the need to reflect on prior thinking.

Investigation of design principle (3)

In accordance with the third intervention design principle, the intervention should ideally cause students to productively engage in analytic thinking on the target question such that they are able to generate a productive model and give a correct response. One would anticipate that an intervention designed to support process 2 engagement in this way would result in students' increased accuracy on the target question after completing the intervention.

However, under the framework of dual-process theories, it is also important to investigate the mechanism responsible for any improvement in performance on the target question. In other words, are students performing better because the intervention questions strengthened their

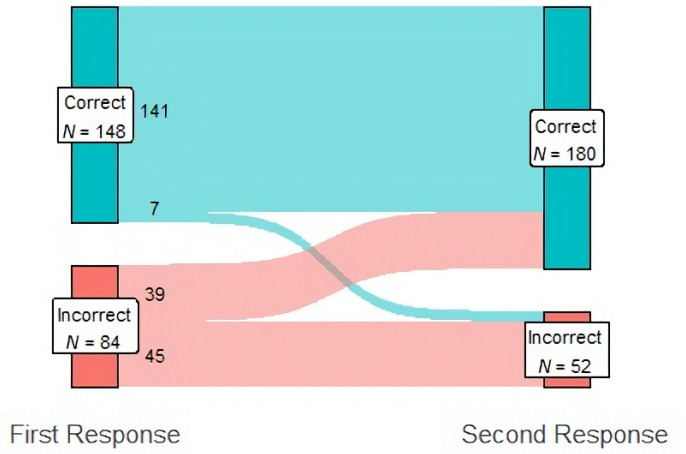


Figure 4.15: Sankey diagram depicting students' responses to the first and second instances of the target question on the Qtr 3.2 question sequence.

conceptual understanding or because it motivated students who already had the relevant content knowledge to effectively reason on the path of sustained effort? If the latter is true, one would expect the intervention to benefit content-proficient students more than noncontent-proficient students.

Analyses testing the above two predictions are discussed below.

More students increase accuracy on the target question than decrease accuracy

In order to evaluate whether accuracy on the target question improved after completing the intervention questions, a McNemar test was conducted on the data shown in Table 4.9 and visualized in Fig. 4.15. The results demonstrate a large effect size that there are more students who changed their answer from incorrect to correct on the target question as opposed to the inverse [$p = 1.83 \times 10^{-6}$, Cohen's g effect size $= 0.35$ (large)]. This analysis supports the notion that the intervention caused students to improve on the target question.

Table 4.9: Contingency table of Qtr 3.2 student responses to the first and second instances of the target question.

	Target second response correct	Target second response incorrect	Total
Target first response correct	141	7	148
Target first response incorrect	39	45	84
Total	180	52	232

Intervention has similar benefit for content-proficient and noncontent-proficient students While the intervention benefited all students, one would anticipate that content-proficient students benefit to a greater degree than noncontent-proficient students. This is because the intervention is targeted towards students who are making reasoning errors and is intended to support analytic thinking without developing mindware. Given the intervention, students who have the appropriate content knowledge should be able to overcome incorrect process 1 models by engaging in more productive process 2 thinking. On the other hand, students without the content knowledge would be unable to produce an accurate model and so no amount of analytic thinking would allow them to arrive at a correct conclusion.

The data set from Qtr 3.2, when divided into content-proficient and noncontent-proficient pools, is too small to conduct sufficiently comparable McNemar tests of students' responses to both instances of the target question due to a discordant cell containing a value of zero. Therefore, a chi-square test was conducted comparing content-proficient and noncontent-proficient student performance on the second instance of the target question for those who incorrectly answered it the first time (see data used in Table 4.10). Sankey diagrams of student responses to both instances of the target question split into content-proficient and noncontent-proficient pools of students are provided in Fig. 4.16. One would expect that

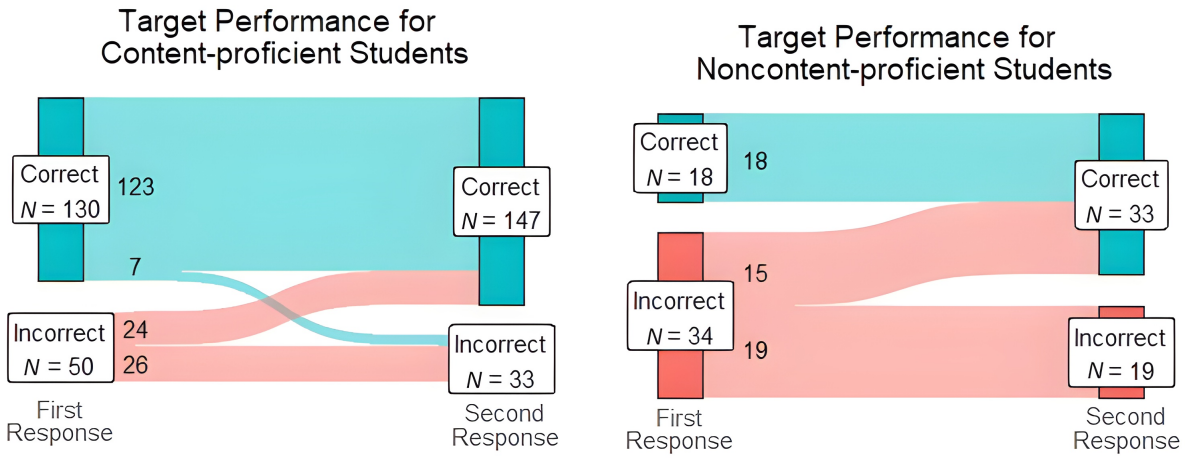


Figure 4.16: Sankey diagrams depicting student responses to the first and second instances of the target question on the Qtr 3.2 question sequence grouped by content proficiency.

content-proficient students would be more likely to switch to the correct answer than those lacking the relevant conceptual understanding. However, there is no difference between these two groups ($p = 0.90$).

There is no evidence to suggest that the intervention is preferentially helping content-proficient students. While it appears that the intervention does benefit content-proficient students, it is also helping noncontent-proficient students improve performance to a similar degree.

One possible reason for this unexpected result is that the screening question may not be properly categorizing students by content proficiency. Maybe the screening question does not assess a complete set of mindware or the strength of mindware needed for the target question and so there could be a pool of students who answer the screening question correctly but lack the necessary conceptual understanding to correctly respond to the target question. In this case their incorrect response to the target question is not a matter of relying on process 1 thinking without engaging in cognitive reflection. Instead, they simply did not

Table 4.10: Contingency table comparing content-proficient and noncontent-proficient Qtr 3.2 student responses to the second instance of the target question for those who incorrectly answered the first instance of the target question.

	Target second response correct	Target second response incorrect	Total
Screening correct	24	26	50
Screening incorrect	15	19	34
Total	39	45	84

have the mindware to be able to answer the target question correctly. These students would be less likely to benefit from a DPTs-based intervention. In this scenario, the group of students identified as content-proficient would contain false positives, diluting any potential difference in performance on the second instance of the target question between groups of students with differing accuracy on the screening question.

Another possibility is that the analytic support questions are not only encouraging process 2 reasoning but also improving students' conceptual understanding by teaching content knowledge. In this case, the intervention would allow noncontent-proficient students to access relevant information needed to succeed on the target question, placing all students on a level playing field when responding to the second instance of the target question. In other words, student performance on the target question after encountering the intervention would be similar.

Alternatively, there is a chance that with a larger sample size, a McNemar test comparing the number of students who increased accuracy to those who decreased accuracy on the target question would show a greater effect-size difference for content-proficient students than noncontent-proficient students as expected.

Let us consider how this could be true even if the same chi-square test used earlier in this

section were performed on a larger data set and still produced null results. Consider the Sankey diagrams shown in Fig.'s 4.16a and 4.16b. When comparing content-proficient and noncontent-proficient groups of students using Mcnemar tests, it is investigating how large of a difference there is within each diagram between the two groups of students whose accuracy on the target question switches (represented by the overlapping flows). When running a chi-square test comparing content-proficient and noncontent-proficient student performance on the second instance of the target question for those who incorrectly responded to the first instance of the target question, it is comparing the distributions of the flows stemming from the bottom left node on each Sankey diagram (in peach).

It is possible for the bottom flows to be comparable on each diagram but the overlapping flows to differ with varying effect sizes on the two diagrams. It is apparent, especially for the noncontent-proficient pool, that a small sample size is inhibiting the amount of information gained from these analyses.

Overall, it appears that the intervention is responsible for generally promoting improved accuracy on the target question for all students. However, it seems that all students benefit to the same extent regardless of content proficiency. Therefore, while the intervention is beneficial, it is unclear to what extent the intervention aligns with design principle (3), namely that the intervention produces improvement by motivating students to engage in productive process 2 thinking. Even if the possibilities listed above are confounding the data, it still seems plausible that the intervention is functioning to support students' analytic thinking; it may just be promoting conceptual understanding as well and/or the analyses prove limited due to other factors associated with the data used.

Performance on the intervention

In addition to validating the intervention based on the DPTs-based design principles, I sought to understand if revisiting and performance on the second instance of the target question were related to student performance on the analytic support questions. One might expect

that students who correctly follow the reasoning of fictional students 1 and 2 would have the greatest benefit from the intervention since they were able to successfully apply Newton's second law on the target question.

Results show that there were associations between success on the analytic support questions and a greater likelihood of choosing to revisit thinking on the target question as well as improved performance on the target question. Additional analyses related to the cognitive reflection test are also presented, which show that the intervention has the potential to benefit all students regardless of their natural tendency towards cognitive reflection. A discussion of the results follows.

Analytic support performance is associated with revisiting For those who answered the target question incorrectly, those who answered both of the analytic support questions correctly were more likely to want to revisit [$p = 0.019$, Cramér's V effect size $= 0.26$ (small)]. The data is presented in Table 4.11.

The risk ratio below quantifies this likelihood. Note that this only includes students who answered the first instance of the target question incorrectly.

$$\begin{aligned}\text{Risk Ratio} &= \frac{\text{(Wanted to revisit \& analytic support correct)/Analytic support correct}}{\text{(Wanted to revisit \& analytic support incorrect)/Analytic support incorrect}} \\ &= \frac{47/56}{17/28} = \frac{0.84}{0.61} = 1.38\end{aligned} \tag{4.5}$$

For students who initially failed to correctly respond to the target question, those who answered the analytic support questions correctly were 1.38 times more likely to consider revisiting their thinking than those who answered one or both analytic support questions incorrectly.

Table 4.11: Contingency table displaying Qtr 3.2 data for students who incorrectly responded to the target question comparing their performance on the analytic support questions to their desire to revisit thinking on the target question.

	Wanted to revisit	Did not want to revisit	Total
Analytic support correct	47	9	56
Analytic support incorrect	17	11	28
Total	64	20	84

This gives credence to the notion that the analytic support questions were responsible for prompting revisiting. However, the analytic support was more beneficial at encouraging students to consider reflective thinking when they successfully navigated the analytic support questions. This somewhat limited the effectiveness of the intervention to those who were able to answer the analytic support questions correctly.

Since the intervention was designed to prompt cognitive reflection, and success on the analytic support questions was advantageous for reaping the benefits of this prompting, one would hope that success on the analytic support questions was not associated with students' cognitive reflection skills. In other words, the analytic support questions should benefit all students regardless of their tendency towards cognitive reflection.

To test if there was any association between analytic support question performance and cognitive reflection, a chi-square test was conducted using the data shown in Table 4.12. This test showed no significant difference in analytic support question performance between student groups based on CRT score level ($p = 0.58$).

This suggests that while the analytic support questions caused students to revisit, particularly if they answered the analytic support questions questions correctly, it was not because they already had a tendency toward cognitive reflection. Both those with low and high CRT

Table 4.12: Contingency table for students' CRT score level and analytic support question performance.

	Low CRT (0-1)	High CRT (2-3)	Total
Analytic support correct	36	156	192
Analytic support incorrect	7	33	40
Total	43	189	232

Table 4.13: Contingency table for students' CRT score level and decision to revisit thinking on the target question for those who incorrectly answered the first instance of the target question.

	Wanted to revisit	Did not want to revisit	Total
Low CRT (0-1)	19	0	19
High CRT (2-3)	45	20	65
Total	54	20	74

scores were equally likely to succeed on the intervention questions and have access to the benefit of an increased likelihood of choosing to revisit their thinking on the target question.

Additionally, for students who answered the original target question incorrectly, those with low CRT scores were observed to have an increased inclination to choose to revisit their thinking after seeing the analytic support compared to those with high CRT scores. A Fisher's exact test was used to determine this difference using the data presented in Table 4.13 [$p = 0.004$, Cramér's V effect size = 0.30 (medium)]. (This test was used instead of a chi-square test since one of the cells in the contingency table had a frequency of zero.)

The following risk ratio for students who responded incorrectly on the target question provides more insight into this difference:

$$\text{Risk Ratio} = \frac{\text{(Wanted to revisit \& low CRT)/Low CRT}}{\text{(Wanted to revisit \& high CRT)/High CRT}} = \frac{19/19}{45/65} = \frac{1.00}{0.69} = 1.44 \quad (4.6)$$

It can been seen that for students initially responding incorrectly to the target question, those with low CRT scores were 1.44 times more likely to choose to revisit their thinking on the target question than those with high CRT scores after being exposed to the intervention. This indicates that the intervention activated process 2 thinking by promoting reflection even for those who had a tendency to forgo reflection, and to a greater extent. The intervention therefore functions to serve students who would naturally be less likely to engage in cognitive reflection.

In summary, for students who originally answered the target question incorrectly, success on the analytic support questions was advantageous for an increased likelihood of recognizing the need to reflect on the target question. While the intervention had less benefit for those who incorrectly answered the analytic support questions, this result indicates that the analytic support questions appear to be the mechanism responsible for prompting a desire to revisit. Additionally, since analytic support question performance was not associated with CRT score level, potential benefit from the analytic support was open to those with high and low cognitive reflection skills equally. Even more encouraging is the fact that for students who answered the target question incorrectly, those with low CRT scores were more likely to recognize the need to reflect than those with high CRT scores after seeing the analytic support questions. Therefore, the analytic support questions appear to be helping students recognize the need to reflect, especially for those who lack a natural propensity for reflection.

Analytic support question performance is associated with improved performance on target question Not only did the analytic support questions prompt reflective thinking, it seemed to improve student performance on the target question as well. For those who answered the target question incorrectly, those who responded correctly to the analytic support questions were more likely to succeed on the second instance of the target question. Table 4.14 shows the data used to perform a chi-square test comparison [p

Table 4.14: Contingency table presenting data for students who failed to answer the target question correctly on the Qtr 3.2 administration of the question sequence. Student performance on the analytic support questions is tabulated along with their performance on the second instance of the target question.

	Target second response correct	Target second response incorrect	Total
Analytic support correct	33	23	56
Analytic support incorrect	6	22	28
Total	39	45	84

$= 0.001$, Cramér's V effect size $= 0.35$ (medium)].

A risk ratio calculation for this comparison is shown below for the group of students who originally answered the target question incorrectly:

$$\begin{aligned} \text{Risk Ratio} &= \frac{\text{(Target again corr \& analytic support corr)/Analytic support corr}}{\text{(Target again corr \& analytic support incorr)/Analytic support incorr}} \\ &= \frac{33/56}{6/28} = \frac{0.59}{0.21} = 2.75 \end{aligned} \tag{4.7}$$

For students who failed to respond correctly to the target question the first time, those who responded correctly to the analytic support questions were 2.75 times more likely to improve accuracy on the target question than those who incorrectly responded to the analytic support questions.

Again, it appears that the analytic support is the mechanism responsible for supporting process 2 engagement, in this case, leading to improved student performance on the target question. This is because more benefit is observed for those who answered the analytic support questions correctly. Note, as tested previously in Sec. 4.7.3, that student success on the analytic support questions was not reliant on cognitive reflection skills and so all students

Table 4.15: Contingency table showing student performance on each instance of the target question for those who answered the analytic support questions correctly on the Qtr 3.2 question sequence.

	Target second response correct	Target second response incorrect	Total
Target first response correct	132	4	136
Target first response incorrect	33	23	56
Total	165	27	192

were equally likely to have the potential to improve their accuracy on the target question after encountering the analytic support questions regardless of their propensity to engage in cognitive reflection.

To gain further insight into students' change in accuracy on the target question based on analytic support question performance, two McNemar tests (with Edward's continuity correction) were conducted. For each group of students designated by their success or failure on the analytic support questions, their response patterns on the first and second instances of the target question were analyzed (data shown in Tables 4.15 and 4.16). For those who answered the analytic support questions correctly, there was a difference in the groups of students' whose accuracy on the target question changed, with a greater portion of students increasing accuracy [$p = 4.16 \times 10^{-6}$, Cohen's g effect size = 0.39 (large)]. For students who answered one or both of the analytic support questions incorrectly, there was no difference between those who switched their answer from incorrect to correct on the target question and vice versa ($p = 0.51$).

Consistent with the results above, the intervention preferentially helped students improve accuracy on the target question for those who were successful on the analytic support questions.

Table 4.16: Contingency table with students' response accuracy on the first and second instances to the target question for those who answered one or both of the analytic support questions incorrectly for Qtr 3.2.

	Target second response correct	Target second response incorrect	Total
Target first response correct	9	3	12
Target first response incorrect	6	22	28
Total	15	25	40

Therefore, the intervention seemed to support improvement.

Similar to the analyses conducted above (see Sec.'s 4.7.3 and 4.7.3), I also wanted to evaluate whether increased accuracy was observed for all students regardless of their propensity to mediate incorrect intuitive thoughts with cognitive reflection. The intervention was intended to support process 2 thinking, especially for those who exhibited reasoning errors, and so one would hope that students with low CRT scores were able to override incorrect provisional models on the target question just as well as those with high CRT scores.

A comparison of McNemar tests for groups of students with differing levels of cognitive reflection skills shows that students with low CRT scores improved accuracy on the target question just as much as those with high CRT scores. Data and results can be found in Tables 4.17 and 4.18.

More students who responded correctly to both analytic support questions switched from an incorrect to correct response on the target question than the reverse. This suggests that the analytic support questions supported process 2 thinking enough to allow them to ultimately arrive at a correct response on the target question. While this was dependent on their success on the analytic support questions, there was an equal chance for students to correctly answer

Table 4.17: Contingency table of student performance on each instance of the target question for students with low CRT scores during Qtr 3.2. A McNemar test with Edward's continuity correction indicates a large effect size that students were more likely to increase accuracy on the target question than decrease accuracy [$p = 0.016$, Cohen's g effect size $= 0.36$ (large)].

	Target second response correct	Target second response incorrect	Total
Target first response correct	21	2	23
Target first response incorrect	12	7	19
Total	33	9	42

Table 4.18: Contingency table with target performance before and after the intervention for students scoring high on the CRT during Qtr 3.2. A McNemar test results in a large effect size difference between groups of students whose accuracy on the target question changed, with a greater frequency of increased accuracy [$p = 0.0001$, Cohen's g effect size $= 0.34$ (large)].

	Target second response correct	Target second response incorrect	Total
Target first response correct	120	5	125
Target first response incorrect	27	38	65
Total	147	43	190

the analytic support questions regardless of cognitive reflection skills, and those with high and low CRT scores exhibited successful improvement on the target question to a similar degree.

Discussion of intervention results

With regards to DPTs-based intervention design principle (1), the analytic support questions seemed to cause students who incorrectly answered the target question to recognize the need to reflect on their thinking as anticipated. There were also positive results in terms of design principle (3). Student accuracy on the target question improved after encountering the intervention, likely because the analytic support questions supported their process 2 engagement allowing them to successfully navigate the reasoning path of sustained effort. However, it is possible that the analytic support questions also worked to improve conceptual understanding since there did not seem to be a difference in target improvement for content-proficient and noncontent-proficient groups of students.

It appears that the analytic support questions can be attributed to the benefits described above since students were more likely to benefit after succeeding on the analytic support questions. Even though the effect of the intervention seems to be limited to those who were able to answer the analytic support questions correctly, students benefited from the intervention's proposed ability to prompt recognition of the need to reflect and engage the analytic thinking process regardless of their cognitive reflection skills.

This is encouraging given that the intervention is designed to especially target students who possess content knowledge yet relied on process 1 thinking on the target question, a group of students who one would expect to have a greater concentration of low CRT scores compared to other students.

Of students who originally responded incorrectly to the target question, 67% went on to answer both analytic support questions correctly. If more students were able to answer the analytic support questions correctly or the intervention were changed such that success on

the analytic support questions was not associated with revisiting or improved performance on the target question, it would maximize the benefit of the intervention. Even so, a majority of students had the potential to benefit from the intervention, and it serves as a proof of concept for this type of DPTs-based intervention in line with that described in [30].

4.7.4 Comparison of box-friction and pulse results

Recall that the question sequence created in the box-friction context acted as a proof of concept for the screening-target validation method and DPTs-based reflective intervention strategy developed previously in the context of a pulse on a spring [41, 30]. This section compares results in each context to give insight into the generalizability of these methods.

The results discussed in this chapter related to the box-friction context greatly align with the results from the pulse context. In terms of the screening-target question assumptions, combined results from both contexts support assumptions (1) and (2) with the same level of effect sizes on corresponding chi-square tests and with overlapping confidence intervals for binary logistic regression odds ratios. Note that the pulse sequence included a set of additional questions that were originally intended as analytic support questions, but were later used as a more complete picture of students' conceptual understanding beyond the screening question. These additional questions were analyzed to test the ability of the pulse screening question to identify students with appropriate content knowledge. There were no additional questions administered in the box-friction context, and so there is no comparison that can be made in this regard.

In terms of the intervention, of those who were incorrect on the target question, 66% wanted to revisit their response in the pulse context and 76% wanted to revisit in the box-friction context. A chi-square test indicates that there is no significant difference. Thus, it seems that intervention design principle (1) concerned with raising doubt is satisfied to the same extent in both cases. For students who had the opportunity to reanswer the target question, a McNemar test comparison of those who switched their answer to the target question shows

improved accuracy on the target question in both contexts with large effect sizes. In both cases this fulfills intervention design principle (2) that many students were able engage in productive analytic thinking to ultimately arrive at the correct answer.

While the intervention in each case resulted in improved target performance for all students, the mechanism responsible for improvement may differ. In the pulse context, the intervention was more beneficial for content-proficient students, suggesting that the analytic support questions supported students' analytic thinking without further developing conceptual understanding. However, in the box-friction context, there was no difference in improvement on the target question for content-proficient and noncontent-proficient students. It is possible that the analytic support questions here were not only providing process 2 support but also developing students' content knowledge as well.

When comparing how performance on the analytic support questions was related to students' decision to revisit the target question and their improvement on the target question, the results differed between the two contexts. On the latest iteration of the pulse-on-a-spring sequence, there was no association between accuracy on the analytic support questions and student responses to the revisit question, nor was there an association between analytic support performance and target improvement. The data also showed no associations when only considering those students who answered the target question incorrectly the first time. This suggests that all students had the potential to benefit from the intervention regardless of their accuracy on the analytic support questions. While accuracy was not relevant for students' subsequent reflection and process 2 engagement, it still appears that the intervention was responsible for benefiting students since there was a greater impact for this iteration compared to the previous iteration.

In the box-friction context, results also indicated that the intervention was responsible for encouraging student reflection and analytic thinking. However, in this case it was supported by the fact that performance on the analytic support questions was associated with students' decision to revisit in addition to their improvement on the target question for students who

incorrectly answered the target question the first time.

Lastly, analyses conducted with respect to the comparative benefit of the intervention for students with high and low CRT scores showed similar results in both contexts. In the pulse context, there was no difference in content-proficient students' decision to revisit or their improvement on the target question for students with high or low CRT scores. (Analyses here focused on content-proficient students since these students benefited more from the intervention than noncontent-proficient students.) The results indicate that the intervention was not preferentially helping students based on CRT score. In the box-friction context, students who incorrectly answered the target question the first time and had low CRT scores were more likely to want to revisit the target question than those with high CRT scores. (Here students who answered the target question correctly the first time were not included since they tended to opt against revisiting the target question, and had no need to revisit anyway.) This suggests that the analytic support questions were able to prompt students to consider reflecting, especially those who naturally had a lower propensity towards cognitive reflection. When comparing target improvement for students with high and low CRT scores, there was no difference in the box-friction context. The intervention seemed to encourage analytic thinking for all students regardless of their typical tendency to cognitively reflect.

The similarity of the results in both contexts lends credence to the generalizability of the screening-target validation method and reflective intervention strategy.

4.8 Conclusion

The dual-process theories framework hypothesizes that one can associate two processes with reasoning and decision making. Process 1 is automatic and intuitive while process 2 is deliberate and rule-based. When given a physics task that may contain distracting surface-level features, a student may rely on an incorrect process 1 model and arrive at an incorrect response even when they possess the content knowledge and skills necessary to succeed at the task. Several interventions have been developed and tested in various contexts to leverage

DPTs with the goal of improving student performance on tasks that elicit an incorrect intuitive response.

These interventions usually involve screening and target questions which are meant to distinguish between students' conceptual understanding and reasoning approaches. Then one of two approaches are used to encourage improved student performance on the target question: either student practice designed to automate a correct process 1 response (reflexive) or a reflective strategy designed to support effective engagement of process 2.

In this chapter, a screening and target question involving boxes at rest on rough floors were implemented in tandem with a reflective intervention sequence aimed at supporting process 2 engagement that encouraged students to consider alternative lines of reasoning. The screening and target questions were validated here using prior methods [41], and the intervention - developed using a previous method [30] but applied in a new context - was evaluated. The results were then compared to prior analyses conducted in the context of a pulse on a spring, in which the aforementioned methods were developed.

First, the results suggest that the screening and target questions are functioning as intended under two assumptions: (1) that the screening question is adequately identifying students who possess the content knowledge necessary to succeed on the target question, and (2) that responding to the target question is largely a matter of mediating initially incorrect process 1 ideas with analytic reasoning.

Assumption (1) was supported by the data in that there is a hierarchy in student responses to the screening and target questions, namely that success on the target question necessarily implies success on the screening question. Assumption (2) was observed in that content-proficient student performance on the target question seemed to be related to and predicted by their tendency towards cognitive reflection. While assumption (2) was supported, it was not very strongly backed by the data, which may be due to the nature of the target question itself. It is possible that CRT scores may be only weakly related to target performance if

some students with high CRT scores don't apply their tendency towards cognitive reflection in the box-friction context due to a strong feeling of rightness about their incorrect intuitive ideas. It is also possible that even if students engage in process 2 thinking while reasoning about the target question, the path of sustained effort may not result in the production of a correct model for some students.

Next, the intervention sequence was evaluated for its consistency with a set of design principles formulated under the DPTs framework. Specifically the following two principles were investigated: (1) students who generate an incorrect provisional model on the target question must experience some doubt in this model in order to recognize the need to reflect, and (3) students must have sufficient motivation to engage in effective process 2 thinking to ultimately produce a correct model. (Only principles (1) and (3) were evaluated since principle (2) was not directly measurable.)

The results in Qtr 4 from a comparison between a group given the intervention and a group who were merely asked if they wanted to revisit their answers suggests that the analytic support questions as part of the intervention seemed to encourage students who initially responded incorrectly to the target question to recognize a need to reflect in accordance with principle (1). Thus, merely asking students if they wanted to revisit their thinking on the target question was not as effective at prompting reflection as the analytic support questions. For students incorrectly responding to the target question on the Qtr 3.2 sequence, a majority of them indicated a desire to revisit their thinking and this desire was dependent on their success on the analytic support questions.

The intervention also appeared to prompt improvement on the target question for all students since they were more likely to increase accuracy on the target question than decrease accuracy, a result that was again dependent on student success on the analytic support questions. Since the target question is designed to require content knowledge and cognitive reflection, if the intervention was merely acting as a motivation towards process 2 thinking without teaching content, students without the appropriate content knowledge would not

improve on the target question after encountering the intervention. However, improvement on the target question was independent of students' content-proficiency for those who originally answered the target question incorrectly. Several factors could be contributing to this result. In any case, the intervention appears to cause overall student improvement, which suggests that it is supporting students' analytic thinking in line with principle (3) even if it may also be teaching and/or strengthening mindware as well.

Since the intervention was intended to target students who relied on intuitive thinking on the target question, various analyses were conducted to determine how the intervention was impacting students based on CRT score, which is a measure of the tendency towards miserly processing. Results indicated that performance on the analytic support questions was not associated with students' cognitive reflection scores, and so the intervention had the potential to benefit all students regardless of their general propensity towards cognitive reflection. In terms of promoting recognition of the need to reflect, encouragingly, students who originally answered the target question incorrectly and had low CRT scores benefited more from the analytic support questions than those with high CRT scores. There was no difference in performance shifts on the target question based on CRT score, suggesting that the intervention was just as helpful in encouraging analytic thinking for those who naturally had low or high propensities to do so.

A comparison of these findings with results from the pulse context indicates that the screening, target, and reflective intervention strategy performs similarly. Thus, the question sequence developed in the box-friction context acts as a proof of concept for the DPTs-based methods used to distinguish between students' conceptual understanding and reasoning approaches as well as the intervention strategy intended to support students' analytic thinking in contexts that elicit incorrect intuitive ideas.

4.8.1 Limitations

The data come from a single population comprised of students in an introductory calculus-based physics course at the University of Washington. This may not be a representative sample of introductory physics students more broadly. See Sec. VI A of [41] for demographic information of a similar population at the UW in the quarters prior to this study in comparison with national data.

Additionally, the Qtr 3.2 data set seemed to be too small to provide conclusive results on select analyses performed in this study. In particular, while testing assumption (2) related to the screening and target questions, chi-square tests comparing students' CRT scores to their performance on the screening and target questions were not significant even though there were significant differences using the combined data. The Qtr 3.2 data set was also too small to conduct a proper McNemar test on separate groups of students based on content proficiency, leaving some ambiguity in the mechanism by which the intervention encourages students to improve their performance on the target question, whether through process 2 support only or the added support of conceptual instruction.

There is a small chance that some students in the dataset analyzed here used an AI chatbot to assist with their responses to questions in the sequence. This would produce results that support the above-mentioned assumptions and principles more weakly since some student responses were potentially not based on their own reasoning.

While the purpose of this study was to test prior methods in a new context, there is some interest into the broader impact of the type of intervention discussed here. There is no data to test whether student success on the target question persisted after some time or if the intervention had any effect on students' use of cognitive reflection on far-transfer questions.

Lastly, the intervention is somewhat limited in its ability to reach all students. Students' desire to revisit thinking and students' improved performance on the target question were dependent on correct responses to the analytic support questions, and 67% of students

who initially answered the target question incorrectly succeeded on both analytic support questions. Keep in mind however that for students who initially respond incorrectly to the target question, there was no difference in analytic support question performance between groups of students with different levels of cognitive reflection or content proficiency.

4.8.2 Further research

While the current study more or less successfully validates a screening-target pair and demonstrates the effectiveness of a reflective intervention method in the context of forces and Newton's laws, a larger sample size including a broader population of students would provide a stronger case for the conclusions drawn. This data could also include student responses on additional questions that give insight into the long-term and/or far-transfer impacts of the intervention.

Similar research could also be conducted in different physics contexts to gain broader insight into these DPTs-based methods. Additionally, incorporating this type of intervention in multiple physics contexts throughout a course could be studied to determine more generalized effects on student cognition, particularly evaluating possible long-term and far-transfer impacts. Note that the intervention strategy could also be implemented via various instructional elements of a course beyond online quizzes, including homework, tutorials, in-class work, and others.

4.8.3 Implications for instruction

Awareness of the reasoning process as described by dual-process theories can be beneficial for both instructors and students. Knowledge of DPTs can promote instructor understanding of student reasoning inconsistencies and generate student recognition of the pitfalls in human reasoning common to everyone, potentially reducing feelings of inadequacy.

Since cognitive reflection is an important part of reasoning needed in physics, instructors should not only resolve to promote conceptual understanding but cognitive reflection skills

as well. While not studied yet, repeated incorporation of DPTs-based interventions as part of the classroom culture could lead to increased student engagement of process 2 thinking over time, potentially resulting in improved performance on various questions that commonly elicit incorrect intuitions.

4.9 Acknowledgments

I would like to recognize Paula Heron for her many contributions to the research described in this chapter. I would also like to thank Andrew Boudreaux, Mila Kryjevskaia, Beth Lindsey, Drew Rosen, and MacKenzie Stetzer for their many contributions to this investigation. Additional thanks go out to all the instructors who generously allowed me to modify course materials and collect data in their courses. Peter Shaffer, Nikolai Tolich, Kazumi Tolich, and David Smith were especially helpful in providing me access to course materials and allowing me the opportunity to add questions to online quizzes and collect student responses. I would also like to thank the members of the Physics Education Group at the University of Washington for their feedback during the writing process. This material was 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.

Chapter 5

CONCLUSION

Current educational goals aim to help students develop "21st-century skills." One discipline-specific educational goal that arises from this objective in combination with the physics education literature is to improve student performance by developing their expertise in physics. Part of expertise is not only having conceptual understanding but having effective reasoning skills that can be applied in multiple contexts. This dissertation approaches the goal of helping develop students' reasoning skills from a dual-process-theories perspective. In this work I discuss an effort to characterize students' reasoning inconsistencies and to develop interventions that encourage cognitive reflection. The intervention strategy presented here aims to provide a "quick" set of guided questions that can be repeated in other contexts and incorporated into general course structure without overburdening the course. The intent of such interventions is to improve students' cognitive reflection skills in physics.

Chapter 2 focuses on a methodology involving "screening" and "target" questions. The screening question is intended to identify students with relevant content knowledge, and the target question requires the same content knowledge but tends to elicit an incorrect intuitive idea. Students who demonstrate conceptual understanding on the screening question but fail to respond correctly to the target question may have relied on an incorrect intuition on the target question, which they did not override with analytic thinking. Chapter 2 develops a validation method for assessing screening-target question sequences on their ability to distinguish between students' conceptual understanding and reasoning approaches. The method is tested on a screening-target pair that involves a pulse on a spring. Two main assumptions are evaluated. The first [assumption (1)] is that the screening question identifies

students with the appropriate content knowledge and skills needed for success on the target question. The results show that this assumption was supported for the pulse sequence. This was supported by two different tests. First, the statement, "correctly answering the target question necessarily implies correctly answering the screening question" was supported by student response patterns to the screening and target questions. Second, students who correctly answered the screening question were more likely to succeed on related questions that tested content knowledge needed for the target question more directly compared to students who incorrectly answered the screening question. The second assumption [assumption (2)] is that for students who demonstrate conceptual understanding by correctly answering the screening question, responding to the target question is largely a matter of applying cognitive reflection to override their incorrect intuitive ideas. This assumption was also supported for the pulse-on-a-spring sequence and was demonstrated in two ways. Content-proficient students with a higher propensity for cognitive reflection as measured by a test of cognitive reflection were more likely to correctly respond to the target question than those with a lower propensity for cognitive reflection (although the measure of cognitive reflection as it applies to physics is nuanced). Additionally, an intervention designed to support students' reflection via analytical thinking without developing their conceptual understanding resulted in improved performance on the target question. This chapter establishes a method of analysis for validating whether other screening-target sequences support the two assumptions mentioned above. This can be used as part of the larger effort to improve students' analytical thinking in physics. Namely, screening-target question pairs that support the two assumptions can provide insight into student reasoning, and provide a starting place for developing interventions that help support student reasoning in a variety of contexts.

Chapter 3 takes a deeper look at the intervention described in Chapter 2. It outlines the iterative development of this DPTs-based intervention in the pulse context and assesses the effectiveness of the final version. Students were given a question sequence comprised of screening, target, and intervention questions, concluding with another opportunity to re-

spond to the target question. The intervention was designed to improve student performance on the pulse target question by supporting students' process 2 thinking. It consisted of analytic support questions and an opportunity to revisit thinking on the target question. It appears that the intervention was able to raise doubt in students' unproductive initial ideas, to remind them of relevant content knowledge, and encourage them to engage analytic thinking through the consideration of alternative lines of reasoning. A majority of students who incorrectly answered the target question indicated that they wanted to revisit their thinking on the target question after encountering the intervention. Those who reanswered the target question were observed to have a greater increase in accuracy on the target question than decrease. This chapter therefore provides a method for developing similar interventions in other contexts that can be used as part of curriculum intended to support students' development of cognitive reflection as well as to improve their functional understanding of physics.

Chapter 4 acts as a proof of concept for the screening-target validation methods outlined in Chapter 2 as well as the intervention strategy developed in Chapter 3. Here, a screening and target question in a different context involving boxes at rest on rough surfaces is validated and a related intervention question sequence is assessed. The screening and target questions seemed to align with the two assumptions investigated. The expected screening-target question hierarchy in student responses was observed, and content-proficient students with a tendency towards cognitive reflection were more likely to succeed on the target question. Results also indicated that the intervention was the catalyst for encouraging students to reconsider their thinking on the target question and for promoting effective engagement of process 2. A majority of students who answered the target question incorrectly choose to revisit it, and students showed significant improvement on the second attempt on the target question overall. This chapter demonstrates the applicability of the methods used in Chapters 2 and 3 to a new context and gives credence to the possibility of extending it to other contexts as well to form a collection of DPTs-based question sequences with interventions that serve to help students develop reasoning skills alongside conceptual understanding.

Cognitive reflection, the ability to mediate intuitive thoughts with analytic thinking, is an important skill for physics students to develop, and therefore, instruction should address cognitive reflection. Awareness of human reasoning through the perspective of dual-process theories can be informative for instructors, and one might argue that it has the potential to ease feelings of inadequacy for students. Interventions that encourage effective process 2 thinking as part of an overall curriculum design could potentially aid in students' development of cognitive reflection skills, although further research would need to be conducted.

Future avenues of research could extend this work. The interventions presented here only provide evidence of short-term improvements on particular target questions. Research could be done to determine if these methods are effective at producing long-term or far-transfer impacts on student performance in contexts with salient distracting features that elicit incorrect process 1 models. Additionally, further study of the relationship between the cognitive reflection test and reasoning in physics would provide more insight into observed student response patterns. Student thinking could also be investigated more deeply through eye tracking data, which is a proxy for attentional allocation. Analyzing student attention towards answer choices on a target question could give more insight about their intuitions versus conclusions. These visual patterns obtained from eye tracking data can be used as an additional way to validate screening and target questions, in particular, with regards to assumption (2). Visual attention directed towards various features of the target or analysis support questions could also shed light on what information students are attending to, which could inform intervention development.

Another avenue of research would be to develop similar interventions covering various topics throughout a course and subsequently evaluate if the regular use of such interventions had an effect on students' general propensity towards cognitive reflection. Finally, and more generally in the realm of DPTs research in physics, one could investigate student affect in relation to awareness of the nature of human reasoning as explained by dual-process theories.

BIBLIOGRAPHY

[1] James W. Pellegrino and Margaret L. Hilton, editors. *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press, Washington, D.C., 1 2012.

[2] Lei Bao and Kathleen Koenig. Physics education research for 21st century learning. *Disciplinary and Interdisciplinary Science Education Research*, 1(1):1–12, 12 2019.

[3] Michelene T.H. Chi, Miriam Bassok, Matthew W. Lewis, Peter Reimann, and Robert Glaser. Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cognitive Science*, 13(2):145–182, 4 1989.

[4] Pamela Thibodeau Hardiman, Robert Dufresne, and Jose P Mestre. The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition*, 17(5):627–638, 1989.

[5] Alan H. Schoenfeld and Douglas J. Herrmann. Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5):484–494, 9 1982.

[6] A. Brown. Analogical learning and transfer: What develops? In Stella Vosniadou and Andrew Ortony, editors, *Similarity and analogical reasoning*, pages 369–412. Cambridge University Press, New York, 1989.

[7] D.N. Perkins and Gavriel Salomon. Are Cognitive Skills Context-Bound? *Educational Researcher*, 18(1):16–25, 1989.

[8] Gavriel Salomon and David N. Perkins. Rocky Roads to Transfer: Rethinking Mechanism of a Neglected Phenomenon. *Educational Psychologist*, 24(2):113–142, 1989.

[9] John D. Bransford, Ann L. Brown, and Rodney R. Cocking, editors. *How People Learn: Brain, Mind, Experience, and School*. National Academies Press, Washington D.C., expanded edition, 2000.

[10] Lei Bao, Kathleen Koenig, Yang Xiao, Joseph Fritchman, Shaona Zhou, and Cheng Chen. Theoretical model and quantitative assessment of scientific thinking and reasoning. *Physical Review Physics Education Research*, 18(1):010115, 6 2022.

[11] Jamie Lee Jensen and Anton Lawson. Effects of collaborative group composition and

Inquiry instruction on reasoning gains and Achievement in undergraduate biology. *CBE Life Sciences Education*, 10(1):64–73, 3 2011.

[12] J. Caleb Speirs, Mackenzie R. Stetzer, Beth A. Lindsey, and Mila Kryjevskaia. Exploring and supporting student reasoning in physics by leveraging dual-process theories of reasoning and decision making. *Physical Review Special Topics - Physics Education Research*, 17(2):20137, 2021.

[13] Rachel E. Scherr. Modeling student thinking: An example from special relativity. *American Journal of Physics*, 75(3):272–280, 3 2007.

[14] Andrew F. Heckler. The Role of Automatic, Bottom-Up Processes: In the Ubiquitous Patterns of Incorrect Answers to Science Questions. *Psychology of Learning and Motivation - Advances in Research and Theory*, 55:227–267, 2011.

[15] David Hammer, Andrew Elby, Rachel E Scherr, and Edward F Redish. Resources, Framing, and Transfer. In J. Mestre, editor, *Transfer of Learning from a Modern Multidisciplinary Perspective*, pages 89–119. Information Age Publishing, 2005.

[16] David Hammer. Student resources for learning introductory physics. *American Journal of Physics*, 68(S1):S52–S59, 7 2000.

[17] Edward F Redish. A Theoretical Framework for Physics Education Research: Modeling Student Thinking. In *Proceedings of the International School of Physics "Enrico Fermi"*, pages 1–63, Bologna, Italy, 2004. Italian Physical Society.

[18] Andrew F. Heckler and Thomas M. Scaife. Patterns of response times and response choices to science questions: The influence of relative processing time. *Cognitive Science*, 39(3):496–537, 2015.

[19] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

[20] Keith E. Stanovich. Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking and Reasoning*, 24(4):423–444, 2018.

[21] Shane Frederick. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4):25–42, 2005.

[22] Mila Kryjevskaia, Paula R.L. Heron, and Andrew F. Heckler. Intuitive or rational? Students and experts need to be both. *Physics Today*, 74(8):28–34, 2021.

[23] Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

[24] Mila Kryjevskaia, Mackenzie R. Stetzer, and Nathaniel Grosz. Answer first: Applying the heuristic-analytic theory of reasoning to examine student intuitive thinking in

the context of physics. *Physical Review Special Topics - Physics Education Research*, 10(2):1–12, 2014.

[25] Cody R. Gette and Mila Kryjevskaia. Establishing a relationship between student cognitive reflection skills and performance on physics questions that elicit strong intuitive responses. *Physical Review Special Topics - Physics Education Research*, 15(1):13, 2019.

[26] Cody R. Gette, Mila Kryjevskaia, Mackenzie R. Stetzer, and Paula R.L. Heron. Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy. *Physical Review Special Topics - Physics Education Research*, 14(1):10113, 2018.

[27] Mila Kryjevskaia, MacKenzie R. Stetzer, Beth A. Lindsey, Alistair McInerny, Paula R. L. Heron, and Andrew Boudreaux. Designing research-based instructional materials that leverage dual-process theories of reasoning: Insights from testing one specific, theory-driven intervention. *Physical Review Special Topics - Physics Education Research*, 16(2):20140, 2020.

[28] Beth A Lindsey, Andrew Boudreaux, Drew J Rosen, Mackenzie R Stetzer, and Mila Kryjevskaia. Reinforcing mindware or supporting cognitive reflection : Testing two strategies for addressing a persistent learning challenge in the context of air resistance. *Physical Review Special Topics - Physics Education Research*, 20(2):20116, 2024.

[29] Mila Kryjevskaia, MacKenzie R. Stetzer, and Thanh K. Lê. Failure to Engage: Examining the Impact of Metacognitive Interventions on Persistent Intuitive Reasoning Approaches. In *Proceedings of the 2014 Physics Education Research Conference*, pages 143–146, Minneapolis, MN, 2015.

[30] Kristin Kellar and Paula Heron. An Intervention to help students recognize and resolve reasoning inconsistencies: An application of dual-process theories. *Physical Review Physics Education Research*, in preparation.

[31] Gordon Pennycook, Jonathan A. Fugelsang, and Derek J. Koehler. What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, 80:34–72, 2015.

[32] Andrew Elby. What students' learning of representations tells us about constructivism. *The Journal of Mathematical Behavior*, 19(4):481–502, 2000.

[33] Sílvia Mamede, Ted A.W. Splinter, Tamara van Gog, Remy M.J.P. Rikers, and Henk G. Schmidt. Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes. *BMJ Quality and Safety*, 21(4):295–300, 2012.

[34] Paula R.L. Heron. Student performance on conceptual questions: Does instruction matter? *AIP Conference Proceedings*, 1513(October 2012):174–177, 2013.

[35] Rebecca Rosenblatt and Andrew F. Heckler. Systematic study of student understanding of the relationships between the directions of force, velocity, and acceleration in one dimension. *Physical Review Special Topics - Physics Education Research*, 7(2):1–20, 2011.

[36] Don C. Zhang, Scott Highhouse, and Thaddeus B. Rada. Explaining sex differences on the Cognitive Reflection Test. *Personality and Individual Differences*, 101:425–427, 2016.

[37] Maggie E. Toplak, Richard F. West, and Keith E. Stanovich. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition*, 39(7):1275–1289, 2011.

[38] Anna K. Wood, Ross K. Galloway, and Judy Hardy. Can dual processing theory explain physics students' performance on the Force Concept Inventory? *Physical Review Special Topics - Physics Education Research*, 12(2):1–5, 2016.

[39] Stephen Kanim and Ximena C. Cid. Demographics of physics education research. *Physical Review Physics Education Research*, 16(2):20106, 2020.

[40] National Center for Education Statistics. Fast Facts: Enrollment.

[41] Kristin Kellar and Paula Heron. Distinguishing between students' conceptual understanding and reasoning approaches: An application of dual-process theories. *Physical Review Physics Education Research*, 21(1):010141, 1 2025.

[42] Valerie A. Thompson, Jamie A. Prowse Turner, and Gordon Pennycook. Intuition, reason, and metacognition. *Cognitive Psychology*, 2011.

[43] Jonathon St B T Evans. The Heuristic-Analytic Theory of Reasoning: Extension and Evaluation. *Psychonomic Bulletin & Review*, 13(3):378–395, 2006.

[44] A. DiSessa. Knowledge in Pieces. In G. Forman and P. Pufall, editors, *Constructivism in the Computer Age*, chapter 4, pages 49–70. Lawrence Erlbaum Publishers, New Jersey, 1988.

[45] Hunter G. Close, Luanna S. Gomez, and Paula R. L. Heron. Student understanding of the application of Newton's second law to rotating rigid bodies. *American Journal of Physics*, 81(6):458–470, 2013.

[46] Andrew Boudreaux. A dual process-based teaching intervention for terminal speed [Conference Presentation]. In *AAPT Virtual Summer Meeting*, 7 2021.

[47] Mikayla Mays, MacKenzie R. Stetzer, and Beth A. Lindsey. Supporting student construction of alternative lines of reasoning [Conference Presentation]. In *Proceedings of the 2021 Physics Education Research Conference*, pages 277–282, 2021.

[48] Joss Ives and Jared B. Stang. Using cueing from question pairs to engage students in reflective thinking: An exploratory study. In *Physics Education Research Conference Proceedings*, pages 245–250, 2020.

[49] J. Caleb Speirs, Robyn Leuteritz, Thanh K. Lê, Rose Deng, and Shawn W. Ell. Investigating the efficacy of attending to reflexive cognitive processes in the context of Newton's second law. *Physical Review Physics Education Research*, 19(1):010108, 1 2023.

[50] Maggie E. Toplak, Richard F. West, and Keith E. Stanovich. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, 20(2):147–168, 4 2014.

[51] Lillian C. McDermott, Peter S. Shaffer, and University of Washington Physics Education Group. *Tutorials in introductory physics*. Pearson Learning Solutions, New York, updated prelim. 2nd edition, 2012.

[52] David E. Trowbridge and Lillian C. McDermott. Investigation of student understanding of the concept of velocity in one dimension. *American Journal of Physics*, 48(12):1020–1028, 12 1980.

[53] David E. Trowbridge and Lillian C. McDermott. Investigation of student understanding of the concept of acceleration in one dimension. *American Journal of Physics*, 49(3):242–253, 3 1981.

[54] Ronald A. Lawson and Lillian C. McDermott. Student understanding of the work-energy and impulse-momentum theorems. *American Journal of Physics*, 55(9):811–817, 9 1987.

[55] Lillian C. McDermott and Peter S. Shaffer. Research as a guide for curriculum development: An example from introductory electricity. Part I: Investigation of student understanding. *American Journal of Physics*, 60(11):994–1003, 11 1992.

[56] Lillian C. McDermott, Peter S. Shaffer, and Mark D. Somers. Research as a guide for teaching introductory mechanics: An illustration in the context of the Atwood's machine. *American Journal of Physics*, 62(1):46–55, 1 1994.

[57] N. D. Finkelstein and S. J. Pollock. Replicating and understanding successful innovations: Implementing tutorials in introductory physics. *Physical Review Special Topics - Physics Education Research*, 1(1):010101, 9 2005.

[58] C. Slezak, K. M. Koenig, R. J. Endorf, and G. A. Braun. Investigating the effectiveness of the tutorials in introductory physics in multiple instructional settings. *Physical Review Special Topics - Physics Education Research*, 7(2):020116, 12 2011.

[59] Trevor I. Smith and Michael C. Wittmann. Comparing three methods for teaching

Newton's third law. *Physical Review Special Topics - Physics Education Research*, 3(2):020105, 10 2007.

Appendix A

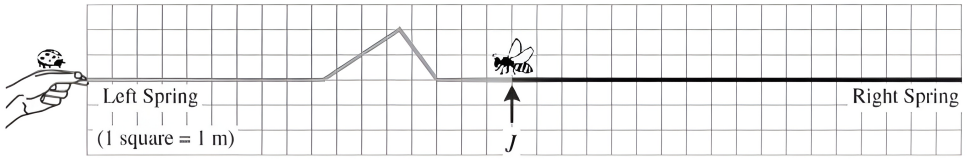
VERSION 3 ADDITIONAL QUESTIONS

The following information concerns the version 3 pulse question sequence described in Chapter 2:

The bug questions used in version 3 of the question sequence are displayed in Fig. A.1 below. The first and third questions were multiple choice while the second and fourth had a short answer format.

The correct answer to the first question is that the time for the bumblebee to move up and down is the same as the time for the ladybug to move, which is the same as the time for the hand to move. The correct answer to the third question is that because the hand motion is unchanged, the time for the bumblebee to move up and down is unchanged in experiment 2.

Recall the original experiment 1 from the previous page in which two different springs are connected at a junction point *J*.



Suppose that in the original experiment (before any changes had been made) that a ladybug had been resting on the student's hand and a bumblebee had been resting at the junction *J*. Both insects are lightweight and have no effect on the pulse.

Recall that in the original experiment, it takes the student's hand a time Δt_0 to quickly move the end of the spring up and down in order to generate the pulse, and the propagation speed of a pulse on the left spring is 1.5 times that on the right spring ($v_L=1.5v_R$).

1. In ***experiment 1***, will the time it takes the bumblebee to move up and down be greater than, less than, or equal to the time for the ladybug to move up and down in this experiment?
2. Explain your response to the previous question.

Just like on the previous page, experiment 2 is nearly identical to experiment 1 except for the ***single change***. Recall that as a result of this change, the width of the generated pulse is doubled. The tension in the spring on the left and the time it takes for the student's hand to move to create the pulse is the same in both experiments. The spring on the right is unchanged.

3. In ***experiment 2*** after the change is made, will the time it takes the bumblebee to move up and down be greater than, less than, or equal to the time for the ladybug to move up and down in this experiment?
4. Explain your response to the previous question.

Figure A.1: Version 3 bug questions.

Appendix B

THREE-BOX FRICTION TARGET QUESTION

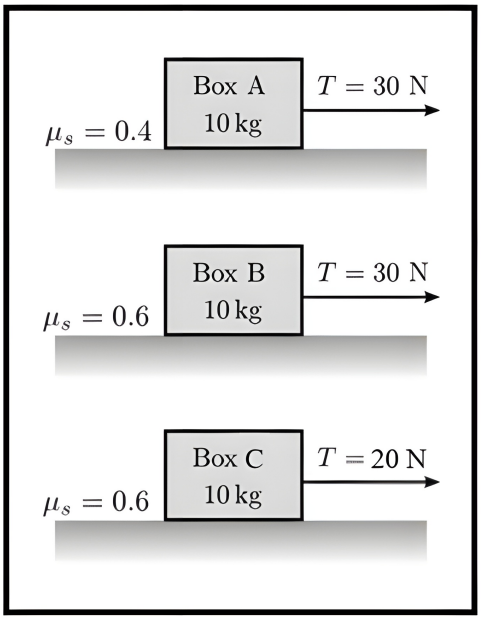
In Chapter 4, a question sequence in the context of forces and Newton's laws is discussed. The Qtr 2 and Qtr 3.2 question sequences in this box-friction context each had two versions, one that included the original two-box target question (see Fig. 4.4), and one with a three-box target question (see Fig. B.1). The three-box version is discussed in more detail here.

On both versions, the setup involves boxes that all have the same mass, experience an applied force, and remain at rest. Boxes A and B in the three-box question correspond to the same boxes in the two-box question. The main difference in the three-box version is the addition of Box C, which has the same coefficient of friction as Box B but less applied force than either of the first two boxes. In this case, students are asked to select the correct ranking of the magnitudes of the friction forces exerted on each of the three boxes and then explain their reasoning.

The correct answer is that the force of friction on Box A is equal to that on Box B, which is greater than the friction force exerted on Box C. Since the boxes are at rest, they must be experiencing balanced forces, and so the frictional forces on each box should be equal in magnitude to the applied forces but directed to the left.

The addition of the third box was intended to provide an opportunity for students to consider how the magnitude of the applied force was related to the friction force in this scenario. On analytic support questions that were presented after the target question, students were asked to consider the arguments of two fictional students and determine how each of these students would respond to the target question. When considering each fictitious student's line of

Suppose there are three boxes, A, B, and C, each on a different floor as shown in the diagram below. The coefficient of static friction between Box A and its floor is 0.4, and there is a horizontal force of 30 N applied to Box A. The coefficient of static friction between Box B and its floor is 0.6, and there is a horizontal force of 30 N applied to Box B. The coefficient of static friction between Box C and its floor is 0.6, and there is a horizontal force of 20 N applied to Box C. $m_A = m_B = m_C = 10$ kg. All three boxes are at rest.



Rank the magnitudes of the friction forces exerted on each box. Explain.

Figure B.1: The three-box target question used on the Qtr 2 and Qtr 3.2 question sequences in the box-friction context.

reasoning, students were not only applying their logic to boxes with differing coefficients of friction but differing applied forces as well.

On subsequent question sequences, the three-box target question version was dropped, and only the original two-box version of the question was used. This decision was made due to the fact that student performance on the two-box question was better than that on the three-box question. Additionally, student responses to the two-box question are already reported in the literature, making the research described in Chapter 4 of this dissertation more comparable.

Appendix C

BOX FRICTION ANALYTIC SUPPORT V1 STUDENT DIALOGUE

As described in Chapter 4, multiple iterations of the box-friction question sequence were administered as part of the development of an intervention. On the Qtr 2 sequence, analytic support v1 questions were given. A fictitious student dialogue was presented (see Fig. C.1) where students 1 and 2 discussed a box (Box B) described in a previous scenario that had a coefficient of friction equal to 0.6 with the floor and an applied force of 30 N directed to the right. The scenario indicated that the box remained at rest. The analytic support v1 questions then asked students how each fictional student would respond to the target question, which appeared earlier in the sequence, and to explain their reasoning. (Note that, in Qtr 2, two versions of the question sequence were given, one with a two-box target question and one with a three-box target question. These questions can be seen in Fig. 4.4 and B.1 respectively.)

Student 1: The force of friction equals μ_s times normal force, so friction is 58.8 newtons on Box B. Box B is at rest because the applied force is less than the frictional force. If the applied force equals 60 newtons, Box B will move.

Student 2: I agree that if the applied force is 60 newtons, Box B will move, but that's because as you push harder on the box the force of static friction increases to match the applied force until static friction can't get any bigger.

Figure C.1: The student dialogue used as part of the analytic support v1 questions on the Qtr 2 box-friction question sequence.