APPLYING THEORETICAL FRAMEWORKS FROM COGNITIVE PSYCHOLOGY TO ASSESS FACULTY PROFESSIONAL DEVELOPMENT AND STUDENT REASONING IN PHYSICS

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Alistair Gilbert McInerny

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Program:
Physics, STEM Education

March 2024

Fargo, North Dakota

North Dakota State University
Graduate School

Title

APPLYING THEORETICAL FRAMEWORKS FROM COGNITIVE PSYCHOLOGY TO ASSESS FACULTY PROFESSIONAL DEVELOPMENT AND STUDENT REASONING IN PHYSICS

By

Alistair Gilbert McInerny

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Mila Kryjevskaia
Chair

Dr. Alexey Leontyev

Dr. Warren Christensen

Dr. Alan Denton

Dr. James Nyachwaya

Approved:

04/07/2024	Dr. Sylvio May
Date	Department Chair

ABSTRACT

Understanding human behavior and reasoning is essential for developing successful instruction. Discipline-based education researchers have examined how students learn, informing the development of successful instructional strategies. Researchers have also identified barriers to the successful implementation of such strategies. This work utilizes two theoretical frameworks from psychology to further examine: 1) efforts to enact instructional change and 2) the effectiveness of instructional approaches to improve students' reasoning in physics. The Theory of Planned Behavior (TPB) is used to assess professional development supporting the successful implementation of evidence-based instructional strategies. The Dual Process Theories of Reasoning and Decision-making (DPToR) are used to model human reasoning and explain persistent inconsistencies in student responses. Guided by the TPB, an assessment instrument was created, validated, and implemented to evaluate instructor's beliefs and intentions about active-learning methodologies. A semi-novel research methodology was also applied to address response-shift bias, a phenomenon common in professional development self-reported assessments. The validation of the instrument and the utility of the retrospective pretest methodology are reported, together with initial assessment results, demonstrating the value of both the TPB and the retrospective pretest in the context of professional development.

The second half of this work discusses inconsistent student reasoning, where students correctly apply conceptual understanding in one context but fail to do so in similar situations. This phenomenon is examined using the Dual Process Theories of Reasoning, which describes reasoning in terms of two processes: a fast, automatic process 1 and a slow, resource-intensive process 2. Process 1 is quick but frequently inaccurate. Process 2 is analytical but time-consuming and effortful. Four reasoning hazards are identified and examined through the lens of

DPToR. Three different types of interventions are implemented to help students develop skills to navigate reasoning hazards: 1) Collaborative exams are used to trigger socially-mediated-metacognition in a high-stakes environment, modeling process 2 activation through group reasoning, 2) a multi-stage guided individual intervention followed by a classroom discussion, and 3) explicit discussion of human reasoning modeled by DPToR. The impacts of these interventions are assessed by comparing results from the treatment (intervention) and controlled (alternative intervention) groups.

ACKNOWLEDGMENTS

This dissertation would not exist without the patience and support of Dr. Mila Kryjevskaia. Thank you for your guidance, leadership, and commitment. You have helped me grow as a researcher, a writer, and a teacher. I would also like to thank Dr. Andrew Boudreaux, who helped me start this journey as an undergraduate.

I would like to thank my committee members, Dr. Warren Christensen, Dr. Alexy Leontyev, Dr. Alan Denton, and Dr. James Nyachwaya, for their feedback and advice throughout this process.

I would like to thank my fellow Physics graduate students for keeping me sane, keeping me company, and helping me along. Thank you, Tim Twohig, Jamie Froeberg, Aaron Fieckert, and Deyan Mihaylov. I would like to thank my friends' partners as well for their hospitality and kindness. Thank you, Elizabith Twohig and Sheila Mihaylov.

Thank you, Sam Brown, Lina Alhalhooly, Reed Peterson, and Wathsala Jayawardana, for always being unreasonably positive and welcoming. And finally thank you to everyone else whose office door frame I leaned against when I needed a break. Thank you for letting me distract you every so often. While I was 1500 miles away from my family, I found friends that made Fargo feel like home.

I would also like to thank every member of the DBER program. Journal Club was my favorite part of graduate school, and I could write a paper on the ways the students and professors in this wonderful group of people showed me kindness and helped me grow. I would specifically like to thank Nathaniel Grosz for taking me under his wing and Rebecca Davidson Reichenbach for so much guidance and camaraderie, as well as everyone I COPUS'd with.

Thank you to every faculty member for making STEM education research fun, welcoming, and thought-provoking.

To the NDSU Physics department, thank you for teaching me and taking me in. Special thanks to Paul Omernik for always helping me quickly and thoroughly with anything and everything, and to John Buncher and Patty Hartsoch for also fitting into the doorframe conversation category.

DEDICATION

This thesis is dedicated to my parents, Diane and Jamie.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike information Criterion
ALS	Active Learning Strategies
ANOVA	Analysis of Variance
ATI	Approaches to Teaching Inventory
Att	Attitudes construct of the Theory of Planned Behavior
BIUALS	Beliefs and Intentions to Use Active Learning Strategies survey
DBER	Discipline Based Education Research
DPToR	Dual Process Theories of Reasoning and Decision-making
EBIP	Evidence Based Instructional Practices
FCI	Force Concept Inventory
FMCE	Force and Motion Conceptual Evaluation
Int	Intention construct of the Theory of Planned Behavior
M	Mean value
NDSU	North Dakota State University
Norms	Subjective norms construct of the Theory of Planned Behavior
PBC	Perceived Behavioral Control construct of Theory of Planned Behavior
PD	Professional Development
PER	Physics Education Research
SD	Standard deviation
TPB	Theory of Planned Behavior
VIF	Variance Inflation Factor

LIST OF APPENDIX FIGURES

1. INTRODUCTION

Over the past few decades, education research has shown that many evidence-based instructional practices (EBIPs) are more effective at improving student conceptual gains than traditional lecture [Freeman et. Al. 2014; Vickrey et al 2015; Theobald et. al. et al 2020], which focuses on the passive transmission of information without actively engaging students in the process of thinking. Research suggests that, if implemented as intended, EBIPs can impact many aspects of student learning [Brewe et al., 2010; Sawtelle et al., 2012; Traxler and Brewe, 2015; Williams et al., 2019; Hazari et al., 2020; McPadden et al., 2020], including improved performance on concept inventories, decreased gaps across genders and ethnicities, greater self-efficacy and retention, improved attitudes towards learning, increased student engagement, improved sense of belonging, and more productive epistemological beliefs.

Despite the proven benefits of active learning methodologies, many factors have slowed their widespread implementation. While the use of EBIPs has been on the rise over the last 20 years, they are still not the norm [Henderson and Dancy, 2008; Borrego et al., 2010; Dancy et al., 2010; Henderson, Dancy, and Niewiadomska-Bugaj, 2012; President's Council of Advisors on Science and Technology (Olson and Riordan), 2012]. Research has identified many obstacles to the successful implementation of instructional innovations (Henderson and Dancy, 2007; Henderson, Dancy, and Niewiadomska-Bugaj 2012, Turpen et al 2016). Significant efforts have been dedicated by researchers in the physics education research (PER) field, as well as across other science, technology, education, and math (STEM) disciplines (generalized as Discipline-Based Education Research or DBER), to understand and alleviate these obstacles. Specifically, researchers have identified the need for targeted professional development (PD) opportunities for college instructors and have developed recommendations for effective PD practices [Henderson, Beach, and Finkelstein, 2011; Henderson, Beach, and Finkelstein, 2012]. At the same time, there is a growing concern that the systematic assessment of PD efforts remains challenging [National Research Council, 2010; American Association for the Advancement of Science, 2019]. As educational change faces resistance, so too do some aspects of student learning. Some patterns of incorrect student responses persist despite decades of efforts in the PER field focused on the development and implementation of EBIPs that target specific student conceptual and reasoning difficulties in physics [Heckler, 2011; Kryjevskaia, Stetzer, and Grosz, 2014, Heckler and Bogdan, 2018; Kryjevskaia, Heron, and Heckler, 2021; Speirs et al., 2019; Lindsey, Stetzer, and Speirs, 2023; Lindsey, Nagel, and Savani, 2017; Gette et al., 2018; Gette and Kryjevskaia, 2019; Kryjevskaia et al., 2015].

This dissertation aims to use theoretical frameworks from cognitive psychology to probe several aspects of teaching and learning that are particularly resistant to change. The Theory of Planned Behavior (TPB) [Ajzen, 1980; Ajzen and Fishbein, 2000; Ajzen and Fishbein, 2011] provides an efficient framework for understanding whether or not a behavior, using EBIPs, for example, will be performed, and the Dual Process Theories of Reasoning and Decision-Making (DPToR) [Kahneman, 2011; Evans, 2006] provides a similarly efficient framework for understanding student reasoning patterns. The Theory of Planned Behavior is used to examine a professional development program designed to support the implementation of instructional innovations on a broader scale, while investigations of student reasoning and the development of instructional strategies to improve student learning are guided by the Dual Process Theories of Reasoning and Decision-Making. The specific investigations discussed here span a wide range of scales and methodologies, from the development and validation of a research instrument designed to assess the effectiveness of professional development programs to specific targeted classroom interventions aimed at improving student learning outcomes.

Chapter 2 of this dissertation discusses the work done as a part of the Gateways-ND professional development program at North Dakota State University (NDSU) [Callens et al., 2019]. This program aimed to change the educational landscape at NDSU by preparing its faculty to implement Evidence-Based Instructional Practices effectively and to support them as they start to implement new practices. The program was designed to help NDSU faculty maximize their instructional effectiveness in STEM courses, resulting in increased student success and engagement in STEM. Gateways ND was expected to produce improvements across various metrics, including completion and pass rates, improved grades, attitudes toward STEM, and study and engagement habits. To achieve this change, the program focused on providing faculty with the necessary training and support to implement EBIPs effectively. A team of NDSU researchers tackled several research questions focused on various aspects of the program [Semanko, 2020; Reichenbach, 2023]. The work described in this dissertation focused on the assessment of the impacts of the Gateways -ND program on behavioral aspects of NDSU faculty that are necessary to enact instructional change. Specifically, a theoretical framework from psychology, the Theory of Planned Behavior was used to develop and validate pre- and post-assessment instrument, the Beliefs and Intentions to Use Active Learning Strategies Survey (BIUALS).

The TPB presents a comprehensive model of behavior that argues that beliefs (separated into individual attitudes, perceived norms, and control beliefs) form the basis of any individual's decision-making to engage in an action or behavior [Azjen and Fishbein, 1980; Azjen and Fishbein, 2011]. While the specific relationship between belief and an action/behavior may vary from person to person or across actions/behaviors, this model is often used to understand which factors are most effective for motivating behavioral changes. This understanding is necessary for informing effective professional development practices that promote instructional and cultural shifts on campus.   This work also applied an innovative methodology, the retrospective pre-test, to mediate response shift bias, a phenomenon common in self-assessments [Howard et al., 1979, Drennan and Hyde, 2008].

Across the Gateways ND research group, data collected from the assessment instrument was used to investigate numerous aspects of PD and its impact on the teaching practices of PD participants. However, this study focuses on two specific aspects of this work: examining evidence for the validity of the BIUALS itself and probing the effectiveness of the retrospective pre-test methodology on mediating response shift bias.

Chapter 3 continues the investigation of assessment practices, however, the focus shifts from the assessment of PD efforts to the assessment of student learning in an introductory physics course. Specifically, the study examines the effectiveness of an emerging student-centered assessment and instructional technique, collaborative exams [Chiu and Kuo, 2010; Cortright et al., 2003; Cooke et al., 2003; Heller et al., 1992; Heller and Hollabaugh, 1992; Jang et al., 2017; Durrant, Pierson, and Allen, 1985; Lusk and Conklin, 2003; Wieman, Rieger, and Heiner, 2014; Lambiote et al., 1987; Stearns, 1996; Yuretich et al., 2001; Gilley and Clarkston, 2014; Knierem, Turner, and Davis, 2015; Zimbardo, Butler, and Wolfe, 2003; Rieger and Heiner, 2014; Efu, 2019]. This investigation was informed, in part, by work in the Gateways professional development program as well as by prior work probing student reasoning in physics. The collaborative exam study was aimed at probing to what degree, if at all, a collaborative exam is effective at helping students resolve inconsistencies between their intuitively appealing but incorrect responses, and formal knowledge of physics gained during instruction.

Prior studies have revealed that students often demonstrate correct physics understanding in one context but fail to apply that same reasoning and knowledge in similar (isomorphic) situations when the new situation tends to elicit intuitively appealing incorrect responses [Gette et al., 2018; Gette and Kryjevskaia, 2019; Kryjevskaia et al., 2014; Kryjevskaia et al., 2015]. This pattern of inconsistencies persists even after the implementation of other EBIPs. The Dual-Process Theories of Reasoning (DPToR) provide a useful lens for examining this phenomenon [Evans and Stanovich, 2013; Thompson, Evans, and Cambell, 2013; Gilovich, Griffin, and Kahneman, 2002]. Specifically, DPToR suggests that inconsistencies in reasoning are due to the nature of human reasoning, which can be modeled as the interaction between two processes. Process 1 is quick and automatic; it immediately suggests a mental model when a reasoner experiences a new situation. Process 2 is slow and deliberate; it is tasked with evaluating the output of Process 1. We hypothesized that, in the presence of the relevant formal knowledge, students' incorrect responses are the results of several factors: 1) salient context features that may trigger a quick, intuitively appealing, and inaccurate response suggested by process 1 [Heckler 2011; Mamede et al 2012; Osman and Stavy 2006], 2) a reasoner’s tendency to accept the output of process 1 without further scrutiny (i.e., cognitive reflection skills are absent or not engaged) [Frederik, 2005; Thompson, 2009; Toplak, West, and Stanovich, 2011; Pennycook et al,. 2016; Campitelli and Gerrans, 2014; Stupple et al., 2013], and 3) even if process 2 is engaged, it is subject to analytical biases (e.g., confirmation bias)[Gilovich, Griffin, and Kahneman, 2002].

We hypothesized that group exams may be an effective instructional technique to engage students in cognitive reflection and address their cognitive biases. We argued that a group discussion during a high-stakes environment (e.g., an exam) might facilitate more productive analytical processing compared to individual work (e.g., limited cognitive reflection) or regular class group work (low-stakes environment). We hope that the high-stakes environment of the exam will motivate students to engage in peer-to-peer discussions, become convinced by peers that their thinking needs to be revised, consider alternatives proposed by peers, identify their mistakes (if present), and correct errors, effectively engaging in socially mediated metacognition [Goos et al., 2002; Vygotsky and Cole, 1978; Shirouzu et al., 2002; Siegel, 2011] . As a result, we also hoped that students would be better able to retain correct knowledge and achieve higher success rates on comparable tasks included in the individual final exam.

Chapter 4 discusses the use of the Dual-Process Theories of Reasoning [Kahneman, 2011; Evans, 2006] to assess a specific student reasoning pattern in the context of Newton's laws and to design an intervention to help students reason more productively. As mentioned above, researchers identified many contexts in which students reason inconsistently [Heckler, 2011; Kryjevskaia, Stetzer, and Grosz, 2014, Heckler and Bogdan, 2018; Kryjevskaia, Heron, and Heckler, 2021; Spiers, 2019; Lindsey, Stetzer and Speirs, 2023; Lindsey, Nagel, and Savani, 2019; Gette et al., 2018; Gette and Kryjevskaia, 2019; Kryjevskaia et al., 2014; Kryjevskaia et al., 2015]. As mentioned above, students often demonstrate correct conceptual knowledge and reasoning in one context and then fail to do the same on conceptually identical questions that tend to elicit intuitively appealing but incorrect responses. This chapter uses DPToR to explore how these inconsistencies in student reasoning can be reduced in the context of Newton's laws.

We also probed the effect of two critical factors on productive reasoning: relevant knowledge and skills (i.e., mindware) [Perkins, 1995], and the tendency toward cognitive reflection [Stanovich, 2009; Frederik 2005], as measured by the Cognitive Reflection Test (CRT)

[Pennycook et al., 2016; Campitelli and Gerrans, 2014; Liberali et al., 2012; Stagnaro, Pennycook, and Rand, 2018; Meyer, Zhou, and Frederick, 2018]. Two conceptually identical questions, one involving a stationary block and the other a magnet on a fridge, were used to demonstrate the phenomenon of intuitively appealing incorrect mental models overshadowing relevant knowledge and skills, resulting in incorrect final responses despite previously demonstrating the necessary skills [Kryjevskaia et al,, 2020]. Most students correctly answer the first question (the Block question) by applying relevant knowledge and reasoning in a context that does not elicit strong, intuitively appealing responses. This question is used to screen for relevant mindware. Roughly half of the students who answered the screening question correctly abandoned that relevant knowledge and provided incorrect, intuitive, responses, cued by salient distracting features of the Magnet task, indicating that the second question served as a target question (eliciting intuitively appealing incorrect responses).

A set of interventions, consisting of question sequences designed with Dual Process Theories in mind, was implemented at three different institutions to guide students through the process of error detection and override. The impact of cognitive reflection skills and mindware was probed at each intervention stage (i.e., before, during, and after intervention).

Chapter 5 discusses a controlled study conducted using DPToR to encourage students to think about their thinking. The study aimed to improve student reasoning consistency by teaching students about their own reasoning processes. The mechanisms of the human brain remain largely unknown despite continued efforts in various research fields to understand them. Considering that even experts do not completely understand the complexities of the human mind, it is reasonable to assume that students are unlikely to be aware of the subtleties and intricacies of their own cognition. This leads us to hypothesize that there are benefits to be gained by teaching students about how they reason.

As mentioned above, many students struggle to apply formal physics knowledge when confronted with situations that have distracting features which prompt intuitive but incorrect responses [Heckler, 2011; Kryjevskaia, Stetzer, and Grosz, 2014, Heckler and Bogdan, 2018; Kryjevskaia, Heron, and Heckler, 2021; Speirs et al., 2019; Lindsey, Stetzer, and Speirs, 2023; Lindsey, Nagel, and Savani, 2017; Gette et al., 2018; Gette and Kryjevskaia, 2019; Kryjevskaia et al., 2014; Kryjevskaia et al., 2015]. This is due to a fundamental aspect of human cognition in which these "intuitively appealing" mental models are tied to an evolutionary need to make quick life-or-death decisions. The human brain is effectively hard-wired to make these quick decisions even though life-or-death situations are uncommon in physics classrooms. Students are often unaware of how this fundamental aspect of their cognition affects their thinking in physics courses, and beyond. We hypothesize that dedicating class time to explicit discussions of the interaction between quick and automatic process 1 and slow and deliberate process 2 can improve student recognition of their own reasoning pathways and improve reasoning. Ideally, explicit knowledge of the duality of human reasoning will encourage students to meaningfully assess their first mental models more frequently.

An intervention was developed to teach students about aspects of their reasoning. In the treatment condition, an instructor first discussed DPToR explicitly, then, over the course of an academic quarter, students were given four different opportunities to answer challenging physics questions that elicited intuitive but incorrect responses. Afterward, students reflected on their responses through the lens of DPToR. To assess the impact of this intervention, a modified intervention protocol was implemented in the same quarter in different course sections taught by the same instructors (i.e., controlled condition). The same set of physics questions was presented to students, and they again reflected on their thinking. However, the instructor did not discuss DPToR or the duality of human reasoning during instruction in these sections, creating a "control" to compare the intervention against.

Two different measures were used to assess the impact of the intervention. The Force and Motion Conceptual Evaluation (FMCE) [Thornton and Sokoloff 1998, Ramlo 2008] was administered before and after instruction to probe the impact on student performance on questions that tend to elicit strong intuitively appealing responses. Student performance on the Cognitive Reflection Test [Frederick 2005; Pennycook et al 2016; Campitelli and Gerrans 2014; Liberali et al 2012; Stagnaro, Pennycook, and Rand 2018; Meyer, Zhou, and Frederick 2018] was also used to gauge whether either mode of intervention engaged student cognitive reflection skills more productively.

In summary, this dissertation presents the findings of several research studies that focus on teaching physics and other STEM subjects, as well as on student reasoning in physics. Two theoretical frameworks derived from cognitive psychology were used as a lens in these studies. These projects varied in scope and objectives. The first study used the Theory of Planned Behavior to examine the professional development of STEM instructors who want to incorporate evidence-based instructional strategies in their courses. The remaining studies focused on the cognitive processes involved in students' thinking in physics courses, using the Dual-Process Theories of Reasoning and Decision-making. The outcomes of the latter studies were used to develop and evaluate instructional strategies designed to improve students' productive reasoning in physics.

2. EXAMINING THE EFFICACY OF A PROFESSIONAL DEVELOPMENT TOOL[1]

Chapter 2 discusses the work done in the context of Gateways-ND, A professional development program at NDSU. This program was developed to teach and provide support for STEM instructors implementing evidence-based instructional strategies in their courses. The Gateways-ND research team examined various aspects of the program. The goal of this particular project was to assess the program using self-reported data from Gateways-ND participants. To achieve this goal, the Theory of Planned Behavior was used to develop an assessment instrument called the Beliefs and Intentions to Use Active Learning Strategies survey. This chapter focuses on examining the validity of this assessment instrument, as well as assessing early data collected with it. Additionally, a novel to STEM Education Research methodology, the retrospective pre-test, was implemented to address a common issue in self-reported data: response-shift bias.

The chapter is split into two subsections. The first focuses on the development and validation of the instrument, the second discusses preliminary use of the instrument for assessing the efficacy of a professional development program at NDSU.

Subsection 2.1 begins by discussing the motivations for assessing the PD programs and the challenges involved in this process. It then goes on to discuss the Theory of Planned

[1] This chapter is largely based on a published paper that was co-authored by Alistair McInerny, Mila Kryjevskaia, and Alexey Leontyev [McInerny, Kryjevskaia, Leontyev, 2021]. Alistair McInerny held primary responsibility for data processing, analysis, and drafting of the paper. Mila Kryjevskaia and Alexey Leontyev provided feedback and guidance through close collaborations and contributed to the editing and revision.

Behavior and the BIUALS' design, as well as the methodology for assessing its validity. The phenomenon of response-shift bias is introduced, and retrospective pre-test methodology is introduced to address it. (Evidence of response-shift bias is discussed in subsection 2.2) Evidence for the validity of the BIUALS is presented and interpreted, covering internal structure, reliability, temporal stability, and relation to known variables.

The second subsection of this chapter provides a concise overview of a preliminary study of BIUALS responses to assess the efficacy of the Gateways-ND program. This assessment was based on the data collected during the program's first three cohorts. The results section highlights patterns and changes in participants' attitudes and beliefs due to their experiences with Gateways ND. Evidence of response shift bias and the relevance of the retrospective pre-test methodology is examined, and arguments are made for the value of the retrospective pretest.

2.1. Development and Validation of the Beliefs and Intentions to Use Active Learning Strategies Survey

2.1.1. Motivation

Despite many well-documented benefits, the use of active learning methodologies in the classroom is still generally overshadowed by traditional lecture [NRC 2012; NRC 2013; Freeman et al 2014; Stains et al 2018]. Research into the diffusion of effective active learning strategies (ALS) revealed that simply making potential adopters aware of pedagogical innovations is not likely to enact changes in instructors' behaviors. Henderson and colleagues argue that effective strategies for facilitating change "are aligned with or seek to change the beliefs of the individuals involved; involve long-term interventions, lasting at least one semester; require understanding a college or university as a complex system and designing a strategy that is compatible with this system." [Henderson et al 2011] Currently, dissemination efforts are shifted toward longer-term professional development programs such as workshops and faculty learning communities [Henderson et al 2011; Khatri et al 2015; Borrego and Henderson 2014; Henderson 2008]. However, assessment of their effectiveness remains particularly challenging.

Ideally, observable changes in behavior should provide evidence for the impacts of a PD program (see section 2.1.2.1). Despite increased attention to the development of teaching observation protocols, practical constraints typically limit their implementations [Stains et al 2018] (e.g., training and paying observers, logistical issues, etc.). Instead, it is typical to rely on self-report measures (e.g., surveys, interviews) to collect data on participants' teaching practices and their perceptions of the effectiveness of PD [Chasteen et al 2016; Chasteen and Chattergoon 2020]. These efforts are critical for moving PD research forward. We argue, however, that it is also valuable to explore a different assessment methodology that uses self-reported data to predict changes in behavior.

Our team drew on PD practices in other fields and psychology research to design, implement, and evaluate an assessment methodology based on the Theory of Planned Behavior, which has been successful at explaining and predicting a range of behaviors, such as smoking and exercising [Ajzen and Fishbein 1980, Ajzen and Fishbein 2011]. The TPB framework was used to develop an instrument that relies on participants' self-reported data to predict their intentions towards implementation of ALSs. Those intentions determine whether this behavior will be performed. Work done by other members of the group did find a mathematical connection between intentions and teaching, as measured using the Classroom Observation Protocol for Undergraduate STEM [Smith et al., 2013; Stains et al., 2018] observations, with cluster analysis providing further evidence of the usefulness of the survey developed here. [Semenko and Ladbury 2020]. The data suggests that instructors with higher post-PD intentions to adopt ALS spend more time using peer instruction techniques (e.g., clicker questions) and/or group work, and less time lecturing. This empirical result could be considered “proof of concept” for the utility of the instrument for PD; however, the study does not discuss evidence for the validity and reliability of measurements, which is the primary goal of this work. We use the framework presented by Arjoon et al. to examine evidence for reliability, temporal stability, internal structure, and relation to other variables [Arjoon, Xu, and Lewis 2013]. We also discuss an important element of our assessment methodology, the retrospective pre-test, which addresses a phenomenon characteristic of self-report measures, response-shift bias, which is not addressed by traditional pre-/post-test methodology [Drennan and Hyde, 2008; Howard et al., 2006].

2.1.2. Design and Methodology

2.1.2.1. Theoretical Framework: The Theory of Planned Behavior

The Theory of Planned Behavior [Ajzen and Fishbein 1980, Ajzen and Fishbein 2011] suggests that an individual’s intent (Int) to perform an action (e.g., implement ALS) can be predicted by a combination of three factors: Attitude toward the behavior (Att), Subjective Norms about the behavior (Norms), and Perceived Behavior Control (PBC) regarding the behavior, explained as follows:

Attitude Toward the Behavior: an individual's evaluation of a specific behavior. This includes the respondent’s beliefs about the outcomes and consequences of the behavior and the degree to which they value or disvalue those outcomes and consequences. In the context of our study, we are specifically interested in participants' attitudes towards “their own use of ALS” in their classrooms. While a participant can have many opinions and beliefs about a concept, the purpose of this factor is not to distinguish those nuances but rather to report their overall effect.

Subjective Norms: the influence of social norms and the opinions of relevant others on the respondent's decision to engage in a behavior. The Subjective norms factor is concerned with the respondent's perception of whether important others (e.g., friends, family, coworkers) would approve or disapprove of the behavior, as well as how motivated they are to comply with these perceived norms. Similar to attitudes, the TPB is not concerned with the respondents' specific perceived norms or whether they are accurate. All that matters is their overall aggregate perception of norms.

Perceived Behavioral Control: the respondent's perception of their ability to perform the behavior in question. PBC is determined by factors such as self-efficacy (one's belief in their capability to perform the behavior) and the presence of external barriers or facilitators that might affect their control over the behavior. Actual skill and barriers are, once again, somewhat irrelevant, as the primary influence of intentions is the self-perception of skill and control.

As mentioned, in this framework, the effects of other characteristics or factors, such as demographic characteristics or experiences, are already accounted for by some combination of the three factors above. For example, suppose a participant experienced being stereotyped by race or gender in educational settings. In that case, they might believe that more inclusive teaching methodologies are important, thus exhibiting a more positive attitude towards ALS. They may also feel less supported in their department, affecting their normative beliefs. If they are a victim of the stereotype threat, they may express doubts about their abilities to create an engaging learning environment, thus underestimating their PBC. In this way, the TPB provides a robust and straightforward model for understanding intentions. Many factors determine an individual's perceptions and beliefs about any given topic, which can provide useful information for specific research questions. But the outcome of those factors (real, perceived, or imagined) are various beliefs, attitudes towards the topic (attitudes), the perception of other's attitudes towards the subject (norms), and an assessment of one's ability to engage with the topic (PBC). These beliefs determine one's intentions to act, and those intentions to act affect behavior.

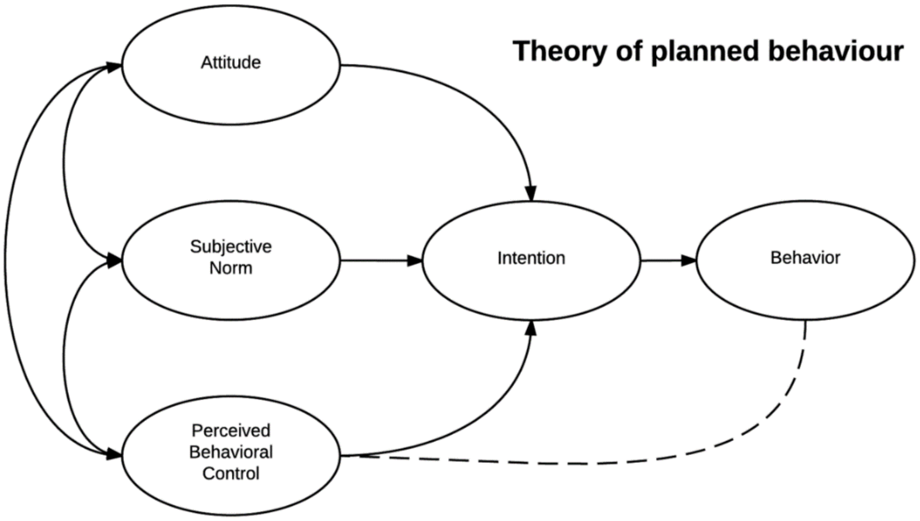


Figure 2.1. Model of the interactions between TPB constructs [Ajzen and Fishbein, 2000].

The specific link between intentions and behavior can be complicated and determining that link can be even more so. An underlying assumption of the TPB is that a positive causal relationship exists between intentions and behavior. We derive the value of our survey from this assumption: positive increases in intentions should correlate with positive increases in behavior. Multiple members of the Gateways ND project attempted to quantify and qualify changes in behavior with varying degrees of success. The following discussion focuses on validating the survey design, followed by some data collection and analysis of survey data to examine the effect of response shift bias and the retrospective pre-test, as well as the impact of Gateways on participants' attitudes.

2.1.2.2. Instrument Design

The development of the instrument (the Beliefs and Intentions to Use Active Learning Strategies Survey, BIUALS) followed standard practices of psychological measurements as prescribed by Ajzen and Fishbein [Ajzen and Fishbein 1980, Ajzen and Fishbein 2011]. A pilot study elicited salient beliefs about behavioral outcomes, normative referents, and control factors. The results were used to create 7-point Likert scale items. Most items utilize response scales ranging from "strongly agree" to "strongly disagree," as shown in Table 2.1. These scales are standard for self-report measures developed in PER [Adams et al., 2006]. Response scales for attitude items use the standard technique, which asks a respondent to evaluate a statement on a set of bipolar adjectives. The adjectives chosen should reveal a person's evaluation or feeling of "favorableness or unfavorabliness" toward the behavior in question. To discourage repetitive answers, reverse scales are employed on several items.

It was also important to develop items that directly correspond to participants' own attitudes toward and intentions to perform a specific action under defined circumstances as opposed to general behavior. As such, all items 1) specify that the statements concern participants' own evaluations (e.g., "for me," "my department") and 2) include a common descriptor of the expected behavior (i.e., the behavior will be performed "in the classroom at some time during the next month."). While seemingly repetitive, these descriptors are necessary. Table 2.1 contains all attitudes, norms, and PBC items, as well as the results of the factor analysis discussed in Section 2.1.3.

Table 2.1. Three and four factor component analysis of BIUALS survey data.

Item	**3 Factors**			**4 Factors**			
	C 1	**C 2**	**C 3**	**C 1**	**C 2**	**C 3**	**C 4**
Attitude 1: “For me, using an ALS in the classroom at some time during the next month is [Extremely Pleasing … Extremely Annoying]”	-	-	-.91	-	-	-.89	-
Attitude 2: “For me, using an ALS in the classroom at some time during the next month is [Extremely Negative … Extremely Positive]”	-	-	-.86	-	-	-.90	-
Attitude 3: “For me, using an ALS in the classroom at some time during the next month is something I [Extremely like … Extremely dislike]”	-	-	-.88	-	-	-.91	-
Attitude 4: “For me, using an ALS in the classroom at some time during the next month is [Extremely worthless … Extremely valuable]”	-	-	-.74	-	-	-.76	-
Attitude 5: “For me, using ALS is in the classroom at some time during the next month is [Extremely punishing … Extremely rewarding]”	-	-	-.88	-	-	-.89	-
Norm 1: “My department chair/head thinks I should use an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.79	-	-	-.72	-	-
Norm 2: “My department chair/head approves of my using an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.81	-	-	-	-	-.85
Norm 3: “My department chair/head wants me to use an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.84	-	-	-.78	-	-
Norm 4: “My department chair/head would support me using an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.77	-	-	-	-	-.90
Norm 5: “My department colleagues think I should use an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.81	-	-	-.91	-	-
Norm 6: “My department colleagues approve of my using an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.85	-	-	-.21	-	-.75
Norm 7: My department colleagues want me to use an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.80	-	-	-.92	-	-
Norm 8: “Most of my department colleagues will use an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.71	-	-	-.59	-	-.22

Table 2.1. Three and four factor component analysis of BIUALS survey data (continued).

Item	3 Factors			4 Factors			
	C 1	C 2	C 3	C 1	C 2	C 3	C 4
Norm 9: “My department colleagues would support me using an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]	-	-.76	-	-	-	-	-.77
PBC 1: “It would be difficult for me to use an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]”	.61	-	-	.54	-	-	-.30
PBC 2: “How much control do you have over whether you use an ALS in the classroom at some time during the next month” [Completely No Control … Complete Control]	.89	-	-	.89	-	-	-
PBC 3: “It is mostly up to me if I use an ALS in the classroom at some time during the next month” [Strongly disagree … Strongly agree]”	.81	-	-	.84	-	-	-
PBC 4: “I feel confident that I could use an ALS” [Strongly disagree … Strongly agree]	.61	-	-.32	.59	-	-.32	-
PBC 5: “I feel confident that I could use an ALS in the classroom at some time during the next month even if I was very busy.” [Strongly disagree … Strongly agree]	.59	-	-	.60	-	-	-
PBC 6: “I feel confident that I could use an ALS in the classroom at some time during the next month even if I was in a bad mood” [Strongly disagree … Strongly agree]	.68	-	-	.67	-	-	-
PBC 7: “I feel confident that I could use an ALS in the classroom at some time during the next month even if the weather was bad” [Strongly disagree … Strongly agree]	.70	-	-	.70	-	-	-
Cronbach’s α	.90	.93	.93	.90	.91	.93	.88

2.1.2.3. Addressing Response Shift Bias, the Retrospective Pre-Test

Traditionally, self-reported data obtained through surveys administered pre- and post-intervention are used to gauge changes resulting from that intervention. Difficulties can arise when using self-report data to assess PD due to an effect known as response shift bias [Howard et al 1979; Drennan and Hyde, 2008; Sprangers & Hoogstraten, 1989]. Effectively, response shift bias is a shift in reference frames that occurs as a result of learning. In the case of PD, response shift bias occurs when a subject’s internal reference frame (used to respond to pre-PD surveys or questionnaires) shifts as a result of the PD itself. The goal of PD is to educate and improve participants' understanding. While growth is ideal, changes in understanding can cause participants to approach post-intervention surveys differently than they did the pre-intervention survey. Pre-post methodologies require that the pre- and post-assessments are comparable or identical, and yet the learning that occurs during an intervention can cause changes in participants' understanding, which can invalidate this assumption. This may be true even for identical surveys because participants’ understanding of the questions themselves can change.

In the case of professional development, participants will hopefully become more experienced with, and knowledgeable about, active learning strategies as a result of the program. While participants will ideally be more prepared to implement active learning in their classrooms, they are also likely to become more aware of the amount of effort and time costs associated with those practices.

For example, instructors with shallow knowledge of active learning, e.g. “I occasionally ask clicker questions in class,” might become less enthusiastic when they learn that proper utilization of clicker questions requires allowing students to spend time discussing their responses (realizing that successful clicker questions will require more class time than they have planned for), or that a large part of the value of clicker questions is their use in formatively addressing student difficulties (again, eating into class time, but also requiring them to sometimes deal with new or unexpected student difficulties). So, while this hypothetical instructor was excited to implement “active learning” before participating in professional development, they leave the program with a more profound but somewhat less enthusiastic understanding, having become aware of the accurate costs and benefits of the practice. A

Traditional pre-post test analysis of this instructor's desire/willingness to use active learning would indicate that they are now less likely to use an active learning strategy as a result of the program. While it is true that this instructor is less excited to use active learning methodologies in their classroom, this is because their prior understanding of active learning methodologies was flawed, and what they were excited to implement in their classroom was not a faithful rendition of the practice. Only considering the pre-post assessment comparison fails to take into account these shifts in the instructor's reference frame and fails to capture that the traditional pre-PD responses were effectively answered using an invalid understanding of the concept (active learning strategies) in question. While the situation described is rhetorical, evidence of response shift bias and its presence among gateways ND participants will be presented later in this chapter.

A response to this phenomenon was pioneered by Howard et al. [1979], who suggested that it might be more accurate for participants to provide self-report data retrospectively. In the previous example, our participant would be asked, after the PD, to think back to their feelings before the PD and answer any self-report questions as they would have answered as if they knew then what they know now. To complete this hypothetical situation from above: before the program, this instructor believes "asking clicker questions" constitutes using active learning and strategies rates their willingness to implement active learning at an 8 out of 10. After the program, this instructor has a much better understanding of the use of clicker questions in regard to active learning. However, they now understand that it takes a fair amount of class time to incorporate them successfully and might require them to spend more time on topics than planned. As such, they now report their willingness to implement active learning as a 7 out of 10. Traditional pre-post analysis of "willingness to implement active learning" would then indicate that the program negatively impacted this participant. While it is true that their willingness to implement active learning decreased, we miss important parts of the picture. If we then ask this instructor to retrospectively score their willingness to implement active learning before the program "knowing what they know now," we might get a response such as the following: "Before the program, I had no intention to spend that much time asking questions, if I'd understood how clicker questions actually worked, I would have been disinclined to use them in the classroom" and rated themselves a 5 out of 10. This retrospective score shows a positive shift from 5 to 7, a positive change despite our negative pre-post scores. Before the PD, the instructor lacked the skill/knowledge to accurately implement clicker questions in their classes, but after the program, while they are still apprehensive about the time and energy such methodologies will take, they have gained experience and resources and are more willing to work them into their instruction. Retrospective pre-tests run somewhat against the grain of traditional pre-post experimental designs despite their ability to show that response shift bias is an issue with such a design. They can provide useful information that a more traditional approach would not. Retrospective pretests have been used in the professional development field as a whole [Hill and Betz, 2005; Lamb & Tschillard, 2005; Lynch, 2002; Nimon & Allen, 2007] to combat this issue but have not made their way into the repertoire of most STEM Ed researchers. In addition to research presented below regarding the stability of retrospective pretest scores, others have also provided evidence for the accuracy and validity of this methodology as well, and even encouraged it instead of traditional pre-tests in situations such as ours [Lamb & Tschillard, 2005; Raidl et al., 2002],

2.1.2.4. Population and Data Collection

The instrument was administered to cohorts of STEM faculty enrolled in a university-wide PD program intended to support faculty adoption of ALS [Callens et al., 2019]. The instrument was given before the start of the program as a pre-test (Pre). It was also administered after the first 2-day workshop as both a post-test (Post-test 1) and as a retrospective pre-test (Retro 1). It was administered again 6 months later, after a semester of FLC and the second 2-day workshop (Post 2 and Retro 2). Retrospective surveys ask participants about beliefs they had before the start of the program (e.g., the first day of the program). This study's multi-year, multi-cohort nature imposed some limitations on the data collection. All participants (N=80) provided responses used in the analysis of the structure and reliability, but only 53 participants responded to both retrospective pre-tests that were matched to probe temporal stability.

2.1.3. Data and Analysis

2.1.3.1. Internal Structure

Analysis of the internal structure was performed to examine the relationships among items. An exploratory factor analysis was done to establish the degree to which the instrument structure is aligned with the theoretical basis for its design, namely items corresponding to the same construct load on the same factor. This analysis is critical for validating the results of the previous “proof of concept” study [Semanko and Ladbury, 2020] since, in the absence of evidence from factor analysis, it would be invalid to average item scores to find a single factor score – an approach used to calculate differences in pre- and post-PD scores to determine change.

Eigenvalues suggest the existence of 4 factors. The scree plot indicates that retaining 3 factors was also reasonable. Extracting 4 factors produced curious results. Two of the four factors loaded exactly as expected (Attitudes and PBC), but the Norms items split into two factors. By inspection, it appears that one subset of these items concerns whether the department wants a participant to implement ALS, while the other subset relates to whether the department will actually support such an action. This result illustrates a dichotomy that often occurs and presents a significant obstacle to instructional innovations: the department recognizes the advantages of ALS but may not necessarily support an initiative to adopt these strategies because of the perceived constraints, such as concerns of reduced content coverage, lack of TA support, poor student evaluations, etc. The detection of such a pattern could indicate a dangerous misalignment between intentions and practices, similar to issues discussed by Henderson et al. [Henderson et al., 2011; Henderson and Dancy, 2007; Henderson and Dancy 2008].

The 4-factor model explains an additional 4% of variance compared to the 3-factor structure. The Norms items, however, do not significantly overlap with other categories, suggesting the viability of the 3-factor structure as well. Extracting 3 variables reveals that all items align with their presumed categories as designed. While it may be advantageous to break Norms items into 2 factors for the reasons above, taken together, they still measure “peers’ attitudes.”

2.1.3.2. Reliability

Reliability was established by exploring a) internal consistency using Cronbach’s alpha and b) temporal stability using a modified approach to test-retest analysis [Drennan and Hyde 2008].

Internal consistency was explored by probing the degree to which participants responding to specific items in a particular way were likely to respond to other items corresponding to the same construct similarly. Items designed to measure a specific construct should correlate with each other; if the correlation is weak or not present, the items should either be removed or redesigned. Analysis revealed that the reliability of each factor is quite high (for both 3- and 4-factor structures). Further investigation into the "reliability if deleted" shows that removing any given item still leaves Cronbach's α above 0.8 for all factors.

For practical considerations, the number of items per construct could be reduced because of the apparent redundancy. While practicality is key for the adoption of the instrument, a significant reduction in the number of items per construct will likely affect the precision of the measurements. Ajzen and Fishbein recommend using at least 3 items per construct. A common rule of thumb for Likert scales is to include at least 4 items per construct to treat average scores as quasi-continuous. If users wish to adopt this instrument as a research tool, retaining 4 items per construct will likely be acceptable. In this case, the practical utility of the instrument may be improved while maintaining precision. If a ceiling effect is suspected, more precise measurements may be needed to detect smaller changes that could be missed via a reduced version.

Temporal stability is a measure of the instruments' reliability over time. While specific experiences are expected to cause participants' responses to change (e.g., instructional interventions), the simple passage of time should not. Typically, temporal stability is established by administering an instrument twice to the same population in a short period of time, eliminating the possibility that any changes in the test-retest measurements, if observed, are due to the intervention. Our study used data from the retrospective pre-tests rather than a more traditional test-retest. Because we were limited to the population of instructors who applied to participate in the PD program, we were not able to identify a period of time during which the participants were not exposed (at least to some degree) to opportunities to think about, consider adopting, or even practice ALS. However, the retrospective pre-test methodology may allow for probing the temporal stability by examining how participants' retrospective self-evaluations of their views of ALS change over time. Specifically, the retrospective pre-test asks participants to re-evaluate beliefs about ALS that they had had prior to PD through the lens of their current understanding. While it is expected that the participants' understanding of ALS will change due to PD, their retrospective self-evaluations should be affected to a much lesser degree, if at all, over the same period of time.

Table 2.2. Correlation coefficients between retrospective scores collected 6 months apart (Retro 1 - Retro 2) and Cohen's *d* of shifts in retrospective and post-test scores collected over the same time period (Retro 1 - Post 2).

Construct	Retro 1 - Retro 2 Correlation Coefficient	Retro 1 - Post 2 Effect Size
Intentions	0.68	0.78
Attitudes	0.55	1.11
Norms	0.74	0.36
PBC	0.66	0.87

Data analysis revealed that scores on retrospective surveys administered 6 months apart remain relatively stable over time. The correlation coefficients between scores are shown in Table 2.2. Paired t-tests did not detect any significant differences in retrospective responses for any construct between Retro 1 and Retro 2. Because these retrospective scores attempt to measure participants' Pre-PD beliefs and are collected 6 months apart, we believe this provides evidence for the stability of retrospective scores.

Paired t-tests applied to retrospective pre-test 1 and post-test 2 indicate desirable and statistically significant shifts in beliefs about and intentions to adopt ALS, as indicated by the effect sizes shown in Table 2.2. (See section 2.1.2.3 for an explanation of why retro-pre-test was used to probe the impact of PD.) We note that the changes in participants' perceptions of norms trail behind changes in other constructs. This result is somewhat expected since changes in community norms are notoriously slow. While participants may start to perceive their community in a more favorable way (perhaps due to support through PD or faculty learning communities), it is likely that the changes in personal beliefs will outpace perceived changes in the norms.

Examining the temporal stability of retrospective pre-test responses provides evidence that the instrument functions as intended, detecting changes due to the PD over time (comparison of Retro 1 and Post 2 scores) while also showing the stability of retrospective self-evaluations measured 6 months apart. This result adds evidence to the quality of inference the BIUALS instrument provides.

2.1.3.3. Relation to Other Variables

Building empirical relationships between the constructs measured by the instrument and other theoretically relevant constructs could provide additional evidence for the validity. The TPB posits that intentions predict behavior. Thus, empirical results that establish the link between the measures of intentions and actual observable behavior constitute strong evidence for the validity of inference (i.e., high intentions will likely lead to desirable behavior). As discussed above, such a link was already established in a prior study. However, further investigations are necessary to explore the functioning of the instrument with different populations and various types of PD.

For completeness and transparency, we note that we also administered the Approaches to Teaching Inventory (ATI) [Trigwell and Prosser, 2004; Trigwell et al 2005] to assess individual’s pedagogical practices along two dimensions: instructor-focused approach and student-focused. The ATI was administered 6 months apart (as a pre-test and after workshop 2) to some cohorts of participants. The instrument did not detect any measurable changes in pre/post- responses. This result is inconsistent with the data from COPUS observations over the same period, which revealed desirable shifts in the adoption of ALSs [Semanko and Ladbury, 2020]. We speculate that ATI items are not sensitive enough (at least for our population and PD). Response shift bias may also be a factor since the ATI was not given in the retrospective format.

2.1.4. Conclusions

We argued for the utility of a new methodology for assessing the effectiveness of PD programs. The TPB informed the development of an instrument to measure intentions, which, in turn, determine behavior. While the link between the instrument's measures of intentions toward the adoption of ALS and the actual behavior was already established empirically in a pilot study, this paper provides evidence for the validity and reliability of these measures. We also advocate for adopting the retrospective pre-test methodology to mitigate the effects of response shift bias in self-report measures.

We speculate that potential users may be concerned about the length of the instrument and the seeming redundancy of some items. We do not dismiss these concerns but remain optimistic that if the adoption of the instrument gains momentum in the community, the community will merge (over time) on a version that balances usability with the quality of measurements.

2.2. Assessing Gateways ND

Professional development is instrumental to the widespread implementation of reformed instructional techniques. To design effective professional development programs, methodologies are needed to evaluate and track the effects of such programs on participant's instructional practices. Many aspects of professional development are critical for its success. This section describes the collection and analysis of data gathered using the BIUALS survey to assess changes in faculty beliefs and intentions to use active learning strategies in their classrooms.

2.2.1. Research Methods

Research population and context. This analysis used data collected from NDSU faculty who volunteered to participate in the Gateways ND professional development program. This cohort consisted of NDSU instructors teaching in a variety of STEM disciplines.

Professional development workshop. A key part of the program is a series of four 2.5-day professional development workshops during a 13-month period over the course of the two-year cohort training cycle. Each cycle starts with a workshop in January, followed by workshops in May (at the end of the spring semester), August (preceding the fall semester), and January. This schedule ensured that the final workshop for each cohort occurs at the same time as the opening workshop for the new cohort, so that the members of the two cohorts can work collaboratively in the Spring. The theme of the first workshop was an introduction to active learning. Discussions of topics were implemented in an interactive, participant-centered format. Workshop topics included principles of active learning, scaffolded instruction, and Backward Design; goals and objectives for a course or a specific lesson; the meaning and scope of assessment; utility of Bloom's taxonomy for evaluating classroom assessment; and implementation of active learning in a SCALE-UP classroom.

The following workshops are designed to focus on the in-depth examination of topics introduced in the first workshop. In particular, the May workshop focuses on gaining expertise in implementing specific active learning strategies in the classroom. The August workshop is intended to provide opportunities for the participants to refine their assessment practices and gain further confidence in transitioning into student-centered instruction. The development of leadership and advocacy among our cohort members was a vital part of the Gateways-ND professional development program. As such, the August workshop draws on the developing expertise of the cohort members in implementing active learning instructional techniques and approaches by inviting these members to lead one of the workshop sessions. The final workshop (January) continues with strategies to develop leadership and focuses on promoting discussion in the classroom. The remaining 11 months of each cohort included the bimonthly meetings of cohort participants, all-campus workshops led by cohort members, and further development of courses that leverage the advantages of ALSs.

Survey data collection and analysis. The pre-workshop survey was administered online a week before the beginning of the first workshop. Post- and retrospective surveys were administered upon completion of the workshop, and a final survey was provided midway through the program after the May workshop. In this analysis, we use data from the first four cohorts of faculty participants at NDSU who volunteered to participate in the Gateways ND professional development program. This cohort consisted of NDSU instructors teaching in a variety of STEM disciplines. Pre-, post-, and retrospective data included in the analysis was gathered from 80 participants. Due to the timing of the data analysis, Post 2 surveys were only completed by 55 members of cohorts 1, 2, and 3. Participants who did not complete the survey at any stage were excluded from this analysis since matched data is required. For each survey category (intentions, attitudes, norms, and PBC), each participant's scores were averaged, and these averages were treated as continuous data due to clusters of at least four questions in each category.

2.2.2. Results

2.2.2.1. Evidence of Response Shift Bias

Multiple one-way Analysis of Variance (ANOVA) were used to determine if statistically significant differences were present between pre-, post-, and retrospective scores for 4 BIUALS categories: Intentions, Attitudes, Norms, and PBC. The results of those tests are reported below, along with the mean scores for each category at each point in time. Post 2 scores were not checked using ANOVA, as using matched data would reduce the number of responses and is not necessary for this part of the analysis. Statistically significant differences were detected in 3 categories: Int, Att, and PBC.

Table 2.3. Mean Likert scores and the statistical significance of the difference between intentions, attitudes, norms, and beliefs before and after the workshop. (N=80)

	Intentions	Attitudes	Norms	PBC
Pre	6.39	5.70	5.25	5.92
Post	6.45	5.83	5.62	6.17
Retro	5.85	5.36	4.93	5.60
F	7.63	4.20	2.15	6.90
P-value	0.0006	0.0161	0.1187	0.012

Because repeated ANOVAs showed statistically significant differences for intentions, attitudes, and PBC, paired t-tests were used to check the statistical significance of differences between time points in each of those three categories. A post hoc $\alpha, \alpha < 0.0055$ was used to determine significance due to the number of tests performed. Differences are reported using effect size (Cohen's d),

Table 2.4. Effect size (Cohen's *d*) of shifts between participant responses at different points in time. Bolded numbers represent p < 0.0055.

	Intentions	Attitudes	PBC
Post-Pre	0.071	0.13	**0.28**
Post-Ret	**0.53**	**0.45**	**0.58**
Pre-Ret	**0.46**	**0.30**	**0.30**

It is important to note that no statistically significant differences were detected in 2 of the 3 categories (i.e., intentions, attitude, and PBC) when using the traditional pre-post test analysis. Despite this, comparing retrospective pretest scores to either pre-scores or post-scores indicates significant differences, which provides some indication of the validity of the retrospective methodology consistent with the predictions of the theoretical framework.

2.2.2.2. Changes as a Result of Professional Development

Pre-, retro-, and post-data only account for differences resulting from the first 2-day workshop. To assess long-term changes, Post 2 data was collected six months later, after participants had experienced a semester of faculty learning communities and their second workshop. Due to the limitations of the pre-scores stemming from the response shift bias, changes are examined by comparing performance on Post 2 and Post, as well as Post 2 and retro surveys.

Table 2.5. Mean Likert scores and the statistical significance of the difference between intentions, attitudes, norms, and beliefs before and after the workshop. (N=80)

	Intentions	Attitudes	Norms	PBC
Post-Post2	0.28	**0.82**	0.22	**0.49**
Retro-Post2	**0.78**	**1.11**	**0.37**	**0.87**

This analysis revealed several indications of Gateways ND's success. Changes between Post and Post2 survey results may serve as a conservative and robust measure of success, representing two well-informed reference frames. A large shift in PBC indicates that the program successfully impacts confidence and self-efficacy. The large impact on attitudes indicates that as participants become more familiar with EBIPs and ALSs, their attitudes recover from the initial downturn predicted (and measured) by response shift bias.

While changes in performance between Post-Post 2 surveys provide a robust measure, it is, as mentioned, a conservative estimate, requiring us to base our success metrics on results obtained after the first professional development. These scores are likely to underrepresent the full shift resulting from an intervention because it ignores any changes that occurred during the first workshop. To take the impact of the first workshop into account, we examined data from the retro survey and compared it to the results of the Post 2 survey. This approach assumes that retrospective scores are a fairly accurate measure of participants' pre-PD beliefs, as argued by several experts in other professional development fields, and also in part due to the temporal stability evidenced earlier. We present results from both Post-Post 2 and Retro-Post 2 comparisons to achieve transparency.

Comparison of Retro and Post 2 scores show substantial shifts in Intentions, Attitudes, and Perceived Behavioral Control, with statistically significant but much smaller (although still medium effect size) shifts in Norms [Sawilowsky, 2009]. This shift between the retrospective- and pre-scores suggests that the Professional Development program had a large impact on participants' attitudes and beliefs about ALS. Again, we note that this interpretation leans on the reliability of the retrospective scores as an accurate measure. However, the comparison of the Post- and Post-2 scores also suggests that, regardless of the accuracy of retrospective results, participants report large and significant changes in their own attitudes and beliefs. While the latter result is weaker, it still posits a significant success in affecting participants' beliefs.

2.2.2.3. Norms

In each of the above cases, participants' scores on the norms questions of the survey are the most resistant to change of the four categories. The result, however, is not surprising. While the workshops aim to change the norms in the community of participants directly involved in the PD program, the impact of this change on the broader institutional communities is not expected on the same short time scale. Indeed, it is expected that the FLCs and interactions with other participants should positively affect participants' norms. It is also recognized that the gateways' participants are from different departments and campus communities, and only a few members of each department participated in the program. As the participants engage with their communities over time, a change in the professional norms may be expected. To determine if this would be the case, aggregate norm scores should be compared across individual cohorts to determine if any shifts occur on a much larger time scale. However, that is not the goal of the current analysis. We note, however, that the preliminary analysis over 4 cohorts presented in Table 2.6 does not reveal any discernable change, which is also expected for the reasons described above.

Table 2.6. Norms by cohort.

	Pre	Retro	Post	Post 2
C1	5.01	4.89	5.0	5.24
C2	5.64	5.33	5.46	5.50
C3	5.058	4.78	5.21	5.57
C4	5.27	4.69	5.38	-

2.2.2.4. Using TPB to Understand Participants' Intentions.

While our BIUALS results have been used above to assess the impact of our professional development program, they can also be used to construct models for the relationships between beliefs and intentions. The TPB posits that some combination of Attitudes, Norms, and Control Beliefs predicts intentions. Here, we use linear regression to create that model.

Table 2.7. Linear regression model coefficients and adjusted R^2. Bolded coefficients represent statistically significant values.

	Pre	Retro	Post	Post 2
Adjusted R^2	0.43	0.78	0.54	0.51
Attitudes	0.17	0.14	-0.04	0.15
Norms	0.05	**0.25**	**0.195**	0.022
PBC	**0.46**	**0.75**	**0.63**	**-0.51**

Retrospective, Post, and Post 2 models explain significant amounts of variance in our data. It is likely that the ceiling effect present in Post and Post-2 data affects their explanatory power, whereas retrospective scores were much less affected by the ceiling effect. The pre-model shows the least explanatory power; however, its validity is compromised (at least in part) by the response shift bias described above. In all 4 models, PBC has the highest impact on intentions in each case, indicating that our participants are largely influenced by their own confidence and perception of control.

2.2.3. Discussion and Conclusion

One of the goals of this work was to incorporate interdisciplinary approaches into a STEM professional development program to enhance our ability to detect changes and assess the effectiveness of professional development efforts. One innovative approach implemented in this work was the retrospective pre-test methodology, used to account for response shift bias, which may be present and often overlooked in STEM Ed professional development research.

Our results indicate that a traditional pre-post approach to assessing a professional development program would effectively yield a null result: professional development produces a minimal effect (mainly on the construct of Control Beliefs after our 2.5-day workshop). The retrospective pre-test yields a more optimistic (and arguably realistic) result. Comparison of pre-scores and retrospective scores revealed significant changes (albeit with small effect sizes) in 3 of the four categories involving self-perception (intentions d = 0.45, Attitudes d = 0.30, PBC d = 0.30). These shifts provide evidence of response shift bias. The goal of the retrospective pre-test is to represent participants' thoughts and beliefs prior to their professional development more accurately. We have shown above that retrospective pretest scores are stable even over a 6-month gap in the survey administration: the participants perceive a significant difference between how they responded to the BIUALs survey before professional development and how they feel they should have responded before the workshop if they knew what they know after the workshop. The results indicate that participants perceive that their pre-test scores overestimate their knowledge and abilities to implement active learning in their own classes. We note that while response shift bias was detected in our data, a replication study is necessary in other professional programs. However, expecting a response shift bias with self-reported data in the context of professional development may help improve PD programs' assessment and development. Given the temporal stability of retrospective scores mentioned above, this methodology provides a more meaningful, consistent, and accurate measure of participants' beliefs and attitudes prior to their engagement with Gateways ND compared to the traditional pre-test methodology. This retrospective methodology could be useful for other professional development programs and other efforts in STEM Education research that involve self-reported data.

Using retrospective scores as a pre-PD measure suggests that the two-day Gateways ND workshop alone significantly impacted participants’ thoughts and beliefs about ALS. Medium effect size shifts were detected in Intentions (d = 0.53), Attitudes (d= 0.45), and Perceived

Behavioral Control (d = 0.58). Compared to the statistically insignificant and minor changes measured by traditional pre-post assessment, we argue that the first Gateways ND workshop had a positive and meaningful impact on its participants. Examining longer-term changes, specifically participants' beliefs during the midway point of the program (Post 2), reveals even more significant changes. Again, a conservative Post-Post 2 methodology underreports the program's total impact, while the Retro-Post 2 comparison provides a more realistic assessment. The conservative approach reports large changes (d= 0.82) for attitudes and medium changes (d=0.49) for PBC. Despite the absence of statistically significant changes in intentions, changes in attitudes and PBC are desirable and detectable even by the more conservative measures that ignore response shift bias. The retrospective methodology that does account for the response shift bias reveals significant changes in 3 of the 4 categories, with moderate but statistically significant changes in the 4th. All contracts show large changes comparing Post 2 scores to retrospective pre-scores (Intentions (d = 0.78), Attitudes (d = 1.11), and PBC (d = 0.87)). This result indicates that Gateways ND has had a significant impact on its participants in terms of attitudes and beliefs. With a moderate and statistically significant change in reported norms, we have evidence that Gateways may also be slowly impacting the educational landscape at NDSU, as was hoped to appear on a larger time scale. These reports of norms are only one small measure of a change in the educational landscape, and more significant claims about the program's long-term impact on that landscape are left to the work of others, as they are beyond the scope of this study.

We explored the relationships between intentions and the other three constructs that predict intentions, as suggested by the theory of planned behavior. Perceived Behavioral Control appears to be the strongest predictor of intentions. This indicates that affecting faculty (self-perception of) confidence and control over their classroom would be the most effective way to increase faculty's intentions to use active learning strategies in their classroom. While attitudes exhibited the most significant shifts throughout the program, their correlation with self-reported intentions was never a significant predictor in our regression models. Shared experiences in the PER community might be confused by this finding: many physics ALS innovators have experienced pushback when introducing those strategies to experimentalist and theorist peers, with the consensus agreeing that attachment to traditional methods was the root cause of this pushback. This may be specific to certain content areas, and as physics instructors are a small minority among gateways ND participants, the shared experience in our field might not translate to institutions as a whole. Examining differences and strengths in departmental cultures could be a fruitful avenue for improving the success of educational reform moving forward.

The impact of norms on intentions also appears insignificant, which contradicts findings from the literature [Henderson and Dancy 2007, Henderson and Dancy 2008]. Further investigation is needed to probe the apparent lack of the relationship. We speculate that, in some departments, faculty are rarely engaged in meaningful conversations about the scholarship of teaching and learning and, as such, may operate largely independently of their peers in the classroom. This is consistent with the overall instructor-centered teaching culture at this institution, which this professional development program hopes to change. Additionally, Gateways ND represents a study of a self-selecting subset of faculty at a specific university, and both the volunteer nature of the program and the specific characteristics of the institution could be responsible for our results. To generalize any findings made using this methodology, replication at other universities is necessary.

To summarize, the Gateways ND program aimed to change the instructor-centered teaching culture at NDSU by providing professional development and support for faculty willing to implement active learning in their classrooms. The theory of planned behavior was used as a framework to implement and evaluate an assessment instrument for probing the impact of that professional development on participants' intentions to implement active learning in their classrooms. The retrospective pre-test methodology was used to evaluate if response-shift bias was present in participants' self-reported data and to provide a new measure (Retro) if necessary. Response shift bias was detected, and traditional methods would have indicated that the workshop had no meaningful effect on faculty beliefs about or intentions to implement active learning. The use of the retrospective pre-test methodology showed that participants were, in fact, positively affected by the professional development. These results underline the utility of the retrospective pre-test methodology for other discipline-based education research projects that rely on self-report data. Any self-reported measures could suffer from response-shift bias, and populations with extremely high or low initial responses are especially at risk.

The implementation of the assessment instrument informed by the theory of planned behavior revealed that Gateways participants experienced significant positive changes in their perceptions of and about active learning, with significant changes in every category using retrospective pre-scores as our pre-intervention measures and significant changes in attitudes and perceived behavioral control even using a more traditional conservative measure. Other studies under the umbrella of the Gateways ND program have demonstrated that these changes in beliefs are, indeed, connected to positive changes in instruction [Semanko and Ladbury 2020].

3. INVESTIGATING A COLLABORATIVE GROUP EXAM AS AN INSTRUCTIONAL TOOL TO ADDRESS STUDENT REASONING DIFFICULTIES THAT REMAIN EVEN AFTER INSTRUCTION[2]

3.1. Introduction

Many research-based instructional materials and techniques positively impact many aspects of student learning, including conceptual understanding and reasoning [Smith et al., 2009; Menekse et al., 2013; McDermott and Redish, 1999; Hsu et al., 2004; Meltzer and Thornton, 2012; Heckler, 2011]. At the same time, a growing body of research suggests that many students continue to reason inconsistently even after targeted instruction designed to address persistent student difficulties [; McDermott and Redish, 1999; Hsu et al., 2004; Meltzer and Thornton, 2012; Heckler, 2011]. In particular, some students demonstrate the necessary conceptual understanding on some physics tasks but fail to do so on isomorphic tasks that require applying the same knowledge and reasoning but also tend to elicit strong, intuitively appealing ideas. [Gette et al., 2018; Gette and Kryjevskaia, 2019, Kryjevskaia et al., 2014; Kryjevskaia et al., 2015].

An overarching goal of our project is to identify factors and instructional circumstances that appear to enhance productive student reasoning in physics. Existing classroom interventions

[2] This chapter is largely based on a published paper that was co-authored by Alistair McInerny, Mila Kryjevskaia [McInerny and Kryjevskaia, 2020]. Alistair McInerny held primary responsibility for data processing, analysis, and drafting of this paper. Mila Kryjevskaia developed the questions in prior studies used to collect data for this project and contributed to the editing and revision of the text.

developed by physics education researchers appear to improve the consistency in student reasoning with various degrees of success. In this project, we probe the efficacy of another (perhaps less conventional) form of instructional intervention, the collaborative group exam, as a strategy to help students identify and resolve inconsistencies in their reasoning that remain even after instruction. Below, we discuss the motivation for, specific aspects of implementation, and this work's implications for teaching.

3.2. Motivation and Theoretical Framework

3.2.1. Dual Process Theories of Reasoning

The Dual Process Theory of Reasoning (DPToR), developed in cognitive psychology, has been used to account for many observed inconsistencies in student reasoning in physics [Evans, 2006; Kahneman, 2011]. The theory suggests that two processes are involved in most reasoning tasks: process 1 is quick, subconscious, and intuitive, while process 2 is slow, logic-based, and deliberate. The most critical aspect of the interactions between the two processes is that process 1 cannot be turned off; we perceive the world around us through the lens of the quick and automatic process 1. Process 2 only intervenes after process 1 has formed an intuition-based mental model. Productive intervention by process 2 requires correct and relevant mindware, a term used to describe the collection of cognitive processes, mental models, problem-solving strategies, and conceptual knowledge that allow reasoners to reason. However, process 2 is often impaired by reasoning biases of its own, and the presence of correct and relevant mindware is occasionally insufficient. For example, individuals tend to look for evidence that supports what they already believe to be true (i.e., confirmation bias) [Nickerson 1998]. To catch a mistake, process 2 must be trained to detect reasoning "red flags." Becoming aware of one's own reasoning and developing the ability to recognize (and act upon) reasoning red flags represents a critical step for developing expertise in physics. These skills are not necessarily physics-dependent, critical thinking skills, problem-solving skills, and even metacognitive skills can apply to thinking and reasoning in any field and should be fostered in all college instruction.

While there are many existing theoretical frameworks for learning and reasoning, we use the Dual Process Theory of Reasoning [Evans 2006, Kahneman 2011] because of its relative simplicity and explanatory power. DPToR posits that there are two processes by which reasoning occurs, and all reasoning pathways can be explained by the interaction between the two processes. As mentioned above, process 1 is a fast, automatic, heuristic process, and process 2 is a slow, deliberate, and analytical process. When confronted with a new situation, process 1 quickly and subconsciously develops a mental model using salient contextual features and readily available prior knowledge. What constitutes a salient feature and what knowledge is readily available depends on the individual reasoner and can be influenced by many factors. Past experiences shape what elements of a situation are noticed and taken into consideration, as well as what knowledge is most easily accessed by the reasoner. Experts in different disciplines (e.g., physics and history) might pay more attention to different elements of the same question based on their experiences or accumulated knowledge such that the heuristics and mental models are also likely to vary significantly. Context and priming can also affect both of these elements. A student in a math class could be primed to focus on different elements of a question than the same student would if the question were presented to them in a physics class. While many different factors affect the formation of the first available mental model suggested by process 1, the mechanisms are the same: the reasoner subconsciously assembles the first impression of a given situation such that it is most plausible to them. The output of this process is often thought of or referred to as a reasoner's intuition.

In many cases, the output of process 1 becomes a final response, and the reasoner moves on. In some cases though, process 2 engages, and the output of process 1 is evaluated. If the first available mental model produced by Process 1 does not satisfy process 2's evaluation the reasoning cycle repeats. Process 1 engages again and produces a new mental model based on different (or reevaluated) knowledge and experiences. Namely, a new mental model may be obtained by consciously or subconsciously modifying the parameters of the next process 1 attempt. A reasoner could focus on different contextual elements of the situation and apply the same heuristics or apply new heuristics to the same contextual elements. While this reasoning process may be iterative, it is not guaranteed to produce a correct response, as process 2 is subject to some of the same biases as process 1 and is also subject to its own biases. The likelihood of process 2 intervention depends on how plausible the reasoner finds their first mental model. This plausibility is strongly related to cognitive miserliness [Fiske and Taylor 1991, Johnson-Laird 2006, Toplak, West, and Stanovich 2011], which is how willing the reasoner is to accept a model as sufficient without analysis.

The functioning of process 2 is what many would refer to as "reasoning." While process 1 is intuitive and subconscious, process 2 is purposeful in its attempt to consider a given situation. Process 2 involves using critical thinking skills, such as troubleshooting, checking limiting cases, and conducting unit analysis, which could serve as effective tools for detecting and correcting mistakes. More general critical thinking skills, such as breaking a problem into smaller parts or systematically approaching the situation, are also part of process 2. While these elements of thinking constitute process 2, they do not define it. Process 2 can produce its own answers or mental models and refine or redirect following Process 1 attempts. Perhaps more importantly, process 2 engagement does not guarantee success. The correct application of accurate mindware is still required even if process 2 is engaged. Factors such as confirmation bias or the lack of another appealing mental model will lead the reasoner to maintain their initial answer.

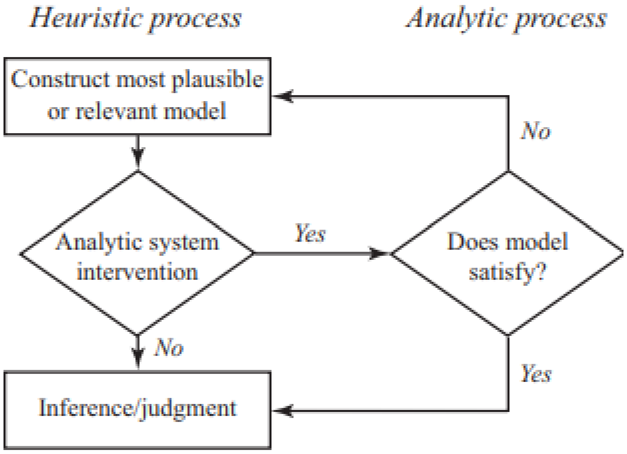


Figure 3.1. An illustration of the heuristic and analytic processes [Evans, 2006]

Many EBIPs target student reasoning difficulties by explicitly engaging students in process 2 thinking and designing instruction that mimics process 2 intervention by asking/forcing students to explain their reasoning or consider alternatives, such as Tutorials in Physics [McDermott and Shaffer 2002]. Classroom interactions may also mimic the questioning and checking of the validity of mental models through student-student and student-instructor dialogs, such as in Think-Pair-Share [Lyman 1981]. However, even under the best classroom circumstances, process 2 might not engage due to low motivation, a tendency toward cognitive miserliness, time constraints, environmental factors, emotional state, etc. [Evans 2006].

Evidence-based instruction often attempts to affect both process 1 and process 2 in several ways. All instruction generally aims to improve students' conceptual knowledge, which is necessary for the effective functioning of both process 1 and process 2, but it is not sufficient.

For example, students who have internalized their conceptual knowledge will access it more easily, leading to correct reasoning from the start. At the same time, these students may also erroneously apply such knowledge inappropriately to situations that require a different set of ideas. Misapplication of ubiquitous physics concepts and ideas often leads to persistent student difficulties that are resistant to classroom interventions. [Kryjevskaia, Stetzer, and Grosz 2014, Heckler 2011, Kryjevskaia, Stetzer, and Heron 2013]. Students often struggle to correctly identify when a specific idea or concept is applicable, frequently reasoning that "since it worked in problem A, and problem B looks like problem A, it must be relevant to problem B too." Despite this, research has shown that even if students reason correctly on one task, they often fail to do so on conceptually identical (isomorphic) situations, not realizing the similarity because "Problem B doesn't look like Problem A." Most troublesome, however, is the research finding these reasoning inconsistencies remain even after evidence-based instruction.

Several contributing factors lead to incorrect or inconsistent student reasoning. One is related to the fluency heuristic, the tendency for reasoners to give more weight to ideas and mental models that come to mind faster [Hertwig et al., 2008, Schooler and Hertwig 2005, Mikula and Heckler 2017]. Similarly, the accessibility of an idea or mental model is also a factor. Accessibility is related to priming and framing, as some ideas are more "close to mind" than others [Kahneman 2003, Morewedge and Kahneman 2010, Heckler and Bogdan 2018, Higgens 1996, Schwarz et al 2003]. A mental model can be more accessible than another if it was learned more recently, or mentioned more recently, or is closely tied (subjectively) with the words or ideas in the question at hand. Famously, many people explain the seasons as the result of the earth being farther away from the sun, resulting in lower temperatures in the winter. Here, the reasoner has applied the heuristic "father away implies less." This heuristic is reliable, heavily used in daily life, and therefore is likely to become fluent. In contrast, the correct explanation using flux is generally much less accessible for most students than many other heuristics, most students do not engage with the concept of flux on any regular basis and are therefore unlikely to reach for it as a resource to solve new problems. While the "farther away implies less" heuristic is accessible, it is at odds with the (somewhat) common knowledge that summer in one hemisphere is winter in the other hemisphere. However, many people do not use this knowledge in explaining seasons because the heuristic "farther away implies less" is used more frequently and works in many cases and the knowledge that seasons are different in different hemispheres is not salient and less readily available for most individuals, highlighting the importance of both accessibility and salience.

Salience refers to the elements of a situation which a reasoner uses to construct their mental models, and many factors can affect which knowledge is salient and which is not. Returning to the example above: the difference in seasons between hemispheres may or may not be recognized as salient by different students. While a student who has never interacted with another hemisphere could easily forget that the season are different across the equator, a student who frequently travels between hemispheres might be more aware of the difference in seasons, and would, therefore, be more likely to reject the initial mental model that "farther from the sun means less heat" due to that heuristic's inability to explain the difference in seasons. The salience of different elements can cause reasoners to approach the problem differently and even cause them to see the problem differently.

It has been observed that even students who possess correct conceptual knowledge may fail to apply that knowledge correctly. For example, when students are asked to rank the forces of a magnet hanging on a fridge and a force is applied upwards on the magnet [Kryjevskaia and

Grosz 2014]. Here the concept of friction is necessary to correctly understand the forces in question. Many students see this situation, and the correct and accessible heuristic “friction opposes the applied force” is triggered, but students incorrectly reason that friction with the fridge must counteract the force applied by the hand. These students fail to recognize the sum of applied forces as a salient feature, which would lead them to correctly apply their heuristic and deduce that the friction must be acting upwards on the magnet. There are many instances in physics in which students fall victim to these intuitively appealing but incorrect first mental models. [Kryjevskaia and Grosz, 2020; Gette et al., 2018; Heron 2017].

The challenges and complexity of developing and correctly applying first mental models demonstrate that training and improving process 1 is not enough to guarantee success, educators must also train students to engage in process 2. Students who engage process 2 are more likely to reason successfully and are more likely to perform well in general [Wood et al., 2016; Gette and Kryjevskaia 2019]. The engagement of process 2 is linked to several aspects of student thinking, such as metacognition and epistemology, but in this study, we focus largely on students’ cognitive reflection abilities. Cognitive reflection can be defined as the ability to moderate intuitive thinking by reasoning more analytically.   As mentioned above, many EBIPs can be thought of as attempts to improve process 2. Critical thinking skills are useful in helping students catch their errors. By asking students to check units, or to check limiting cases, students’ process 2 becomes trained to identify inaccuracies in the initial mental model. Process 2 can also be improved by decreasing cognitive miserliness. This is often done by training students to assess their initial models more often. By asking students to regularly explain their reasoning, and/or by highlighting common errors in that reasoning, students can develop the habit of engaging in process 2. Regarding Metacognition and Epistemologies, if students are more aware of their thinking and more deliberate about their reasoning, they are more likely to engage in process 2 and more likely to do so successfully. While Metacognition and Epistemology are briefly touched on in the final chapter of this paper, they are generally beyond the scope of the studies conducted during my graduate work. To summarize, the activation of process 2 is not enough on its own, students need the relevant conceptual knowledge to answer the question as well as reasoning skills to detect and override errors. As such, a strong conceptual base, or "good mindware," is necessary but not sufficient to reason successfully.

3.2.2. Socially Mediated Metacognition

It has been shown that collaborative group work is effective in engaging students in socially mediated metacognition in which group members share their individual thinking, evaluate ideas, receive feedback, and monitor each other's reasoning [Goos et al., 2002; Vygotsky and Cole, 1978; Siegel, 2011; Chiu and Kuo, 2010]. As such, it is likely that a collaborative work environment may be effective in helping students both identify reasoning red flags and mediate intuitive ideas via analytical reasoning. In recent years, a different form of collaborative group work, the collaborative exam, has gained momentum in the science education community [Cortright et al., 2003; Cooke et al., 2019; Heller et al., 1992; Heller and Hollabaugh, 1992; Jang et al., 2017; Durrant, Piersen, and Allen, 1985; Lusk and Conklin, 2003; Wieman, Rieger, and Heiner, 2014; Lambiotte et al., 1987: Stearns, 1996; Yuretich et al., 2001; Gilley and Clarkson, 2014; Knierim, Turner, and Davis, 2015; Zimbardo, Butler, and Wolfe, 2003; Weiger and Heiner, 2014; Efu, 2019; Elby, 2001; Lindsay, Heron, and Shaffer, 2012]. An emerging body of research suggests that collaborative exams have marked benefits that include improved performance [Lusk and Conklin, 2003; Wieman, Rieger, and Heiner, 2014; Lambiotte et al., 1987: Stearns, 1996; Yuretich et al., 2001; Gilley and Clarkson, 2014; Knierim, Turner, and Davis, 2015; Zimbardo], increased motivation [Weiger and Heiner, 2014], and decreased test anxiety [Wieman, Rieger, and Heiner, 2014]. While collaborative exams appear to be a promising and innovative educational tool, many aspects of their efficacy are still under investigation [Weiger and Heiner, 2014; Efu, 2019]. For instance, instructors often raise concerns that collaborative exams simply promote the propagation of correct answers, do not facilitate individual growth, and are challenging to implement. At the same time, emerging evidence suggests that collaborative exams do often promote individual learning if students are given adequate time to meaningfully examine their responses (even if none of the students were able to arrive at a correct answer on their own) [Durrant, Piersen, and Allen, 1985]. In addition, the impact of group exams is enhanced even further in classrooms where collaborative group work is a norm, thus capitalizing on the alignment between assessment and instruction [Efu, 2019].

As stated above, collaborative group work fosters socially mediated metacognition which may help students learn to recognize reasoning red flags and develop strategies for resolving inconsistencies. The high-stakes environment of an exam setting may boost this effect further by enhancing student motivation to arrive at a correct response with correct reasoning. Because students were graded based on the quality of their reasoning, they may have been particularly motivated to examine their thinking as opposed to just accepting an answer as correct. The latter may happen more frequently during regular classroom instruction.

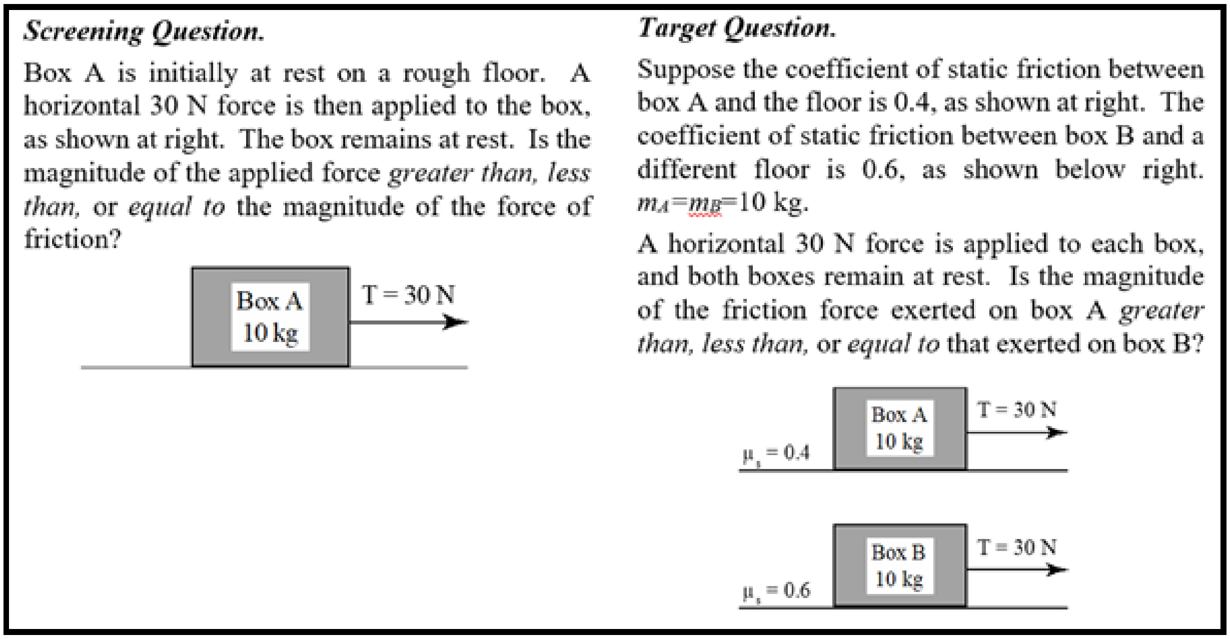
3.3. Methods

This study was conducted in two semesters of an introductory calculus-based mechanics course serving primarily non-physics majors at a mid-size, research-focused land grant university. In both semesters, different instructors implemented active learning techniques such as peer instruction, tutorials, and collaborative group work.

In semester 1, we employed the two-stage exam design, featuring an individual component and a group component, during a 2-hour class period. First, students completed the test individually and submitted their written responses (~60 min); then, following a short break, students were given an opportunity to work in collaborative groups (~40 min). In our study, the collaborative portion included a subset of questions from the individual component, including those questions known to reveal persistent incorrect intuitive responses. Questions from the individual component that are irrelevant to this study were also included. During the collaborative group component, students were allowed to choose their own group partners, but they overwhelmingly stayed within the groups formed during regular classroom instruction. Although students were encouraged to discuss responses to the collaborative component in their groups, they were required to submit their own answers with detailed explanations of their reasoning. In semester 1, 68 students completed individual and group components on 3 midterm exams. Student performance on the group components contributed 20% to their midterm grades. Students did not receive any formal feedback from an instructor during or after the exams. Semester 1 is considered to provide “treatment” conditions.

In semester 2, 48 students completed identical exams individually (no group component was included); however, the instructor conducted a follow-up classroom session dedicated to reviewing solutions to the exam and answering student questions. In both semesters, exam solutions were not posted. Students did not have access to the exam problems to review or “study from” before the final exam. We consider semester 2 to be the “control” condition. We believe that semester 2 represents traditional exam design and therefore serves as a baseline to compare the collaborative exam treatment against. In this study, we intended to probe the efficacy of the collaborative exam as a learning tool rather than an assessment tool, with a specific focus on probing the impact of this intervention on student performance on questions that tend to elicit intuitively appealing (but incorrect) ideas that persist even after classroom instruction.

To probe the level of consistency in student reasoning, we used the screening-target methodology, which employs a pair of isomorphic questions. A screening question probes whether a student possesses the knowledge and skills necessary to analyze a given situation correctly. A target question requires the application of the same formal knowledge and skills but includes surface features that tend to elicit incorrect intuitively appealing ideas.



Screening Question.

Box A is initially at rest on a rough floor. A horizontal 30 N force is then applied to the box, as shown at right. The box remains at rest. Is the magnitude of the applied force *greater than, less than,* or *equal to* the magnitude of the force of friction?

Target Question.

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

A horizontal 30 N force is applied to each box, and both boxes remain at rest. Is the magnitude of the friction force exerted on box A *greater than, less than,* or *equal to* that exerted on box B?

Figure 3.2. Example Target screening question pair.

An example of a screening-target question pair is shown in Figure 3.2. On the screening question, most students correctly recognize that, because box A remains at rest, the net force on the box must be zero; therefore, the force of static friction must be equal in magnitude to the applied 30 N force. The target question requires the application of the same reasoning because both boxes remain at rest while identical 30N horizontal forces act on each box. However, ~25%

of the students who answer the screening question correctly do not use the correct reasoning approach on the target question. Instead, they tend to argue that the magnitude of the friction force on box A is less than that on box B because the coefficient of static friction between the surface and box A is smaller. Through the lens of dual process theory, we argue that the inclusion of the extraneous information on the target question cues an automatic but incorrect response that "higher μ implies higher friction." It appears that students who make this type of error immediately and subconsciously embrace this response as correct, while the mathematical relationship (irrelevant in this case) between kinetic friction f_k and μ_k, $f_k = \mu_k N$, provides further confirmation of the intuitive μ-based response.

The screening-target pair in Figure 3.2 served as one of the 5 pairs of questions included across the individual components of the 3 midterm exams. The other 4 pairs were designed in the context of kinematics graphs [Heckler, 2011], Newton's 3rd law [Elby, 2001], dynamics of circular motion, and work and energy [Lindsay, Heron, and Shaffer, 2012] and are included in appendix A. In both semesters, 1 question pair was included on exam 1 and 2 pairs were included on exams 2 and 3. In semester 1, each target question was also included in the collaborative group exam component. As stated above, students were required to provide in-depth reasoning since their work was graded based on the quality of explanations rather than the correctness of answers.

To assess the effect of both conditions on student performance on questions that tend to elicit incorrect intuitive (rather than correct formal) reasoning, all five target questions were also included on the final exam in both semesters. We recognized that the most significant limitation of this approach was the possibility of students giving memorized responses. However, the decision to include the same set of five target questions on the final exam was made after careful consideration. First, since student intuitively appealing ideas are often cued by specific features of a task, designing new versions of target questions may introduce new variables without necessarily addressing the possibility of a memorized response. As such, we opted for a more parsimonious design. Second, given the persistent nature of student difficulties, we were interested in probing the impacts of the two interventions under the most favorable conditions. Moreover, the results of data analysis discussed in Section 3.3.1 suggest that the memorization of correct responses is not a major factor affecting student performance on the final exam. The time between testing on a midterm exam and re-test on the final exam varied from several months (for midterm 1) to several weeks (for midterm 3).

3.3.1. Results and Discussion

Student individual responses were coded in a binary format with a score of 1 or 0 given to correct or incorrect responses, respectively. Then, each pair of student screening-target responses received one of four possible codes [i,j], with i and j representing performance on the screening and target questions, respectively.

Three approaches to data analysis and interpretation were employed: (1) a course-level analysis involving comparison of all aggregated [i,j] codes pre- and post-treatment in the two conditions, (2) a student-level analysis using a matched pre- and post-treatment data for each student, and (3) a question-level analysis conducted to probe shifts in student performance on each question.

3.3.1.1. Analysis of Performance Pre-Treatment.

A course-level analysis of student performance on the individual components of the three midterms revealed nearly identical results in the two semesters (see Table 3.1) suggesting no difference in the student populations before treatments. Close to half of the total student responses to the five pairs of screening-target questions were correct and consistent (codes [1,1], 42% and 47%): students answered both screening and target questions correctly. Nearly a fifth of all responses were coded as [1,0] revealing inconsistencies in reasoning that suggest that a fraction of the students who are able to apply correct conceptual understanding on screening questions tended not to deploy those correct reasoning approaches on target questions which elicit intuitively appealing responses. More significantly, a third of all correct responses on the screening questions were followed by an inconsistent response on the corresponding target question ([1,0]/([1,1]+[1,0])) suggesting that even in the presence of correct conceptual understanding many students had not yet developed strategies to recognize intuitively appealing, but incorrect, ideas and to override them with correct reasoning acquired during formal physics instruction. Understanding and addressing this type of reasoning errors is an overarching goal of this project.

Table 3.1 Student individual performance on midterms.

Codes	Semester 1 (collaborative group exam condition)	Semester 2 (instructor-led exam review condition)
[1,1]	42%	47%
[1,0]	20%	17%
[0,0]	27%	25%
[0,1]	11%	11%

Approximately a quarter of all responses revealed that students had not developed a basic conceptual understanding as evident by the incorrect responses to both screening and target questions (codes [0,0]).

A small fraction of codes [0,1] demonstrated inconsistencies in student reasoning; however, further research is needed to pinpoint the sources of this error (student carelessness on screening questions, guessing on target questions, etc.). The Sankey diagrams in Figure 3.2 provide some help with data visualization: the vertical bars on the left side of each diagram represent prevalence of specific codes on the midterms (also included in Table 3.1). The bars on the right side represent student performance on the final exam.

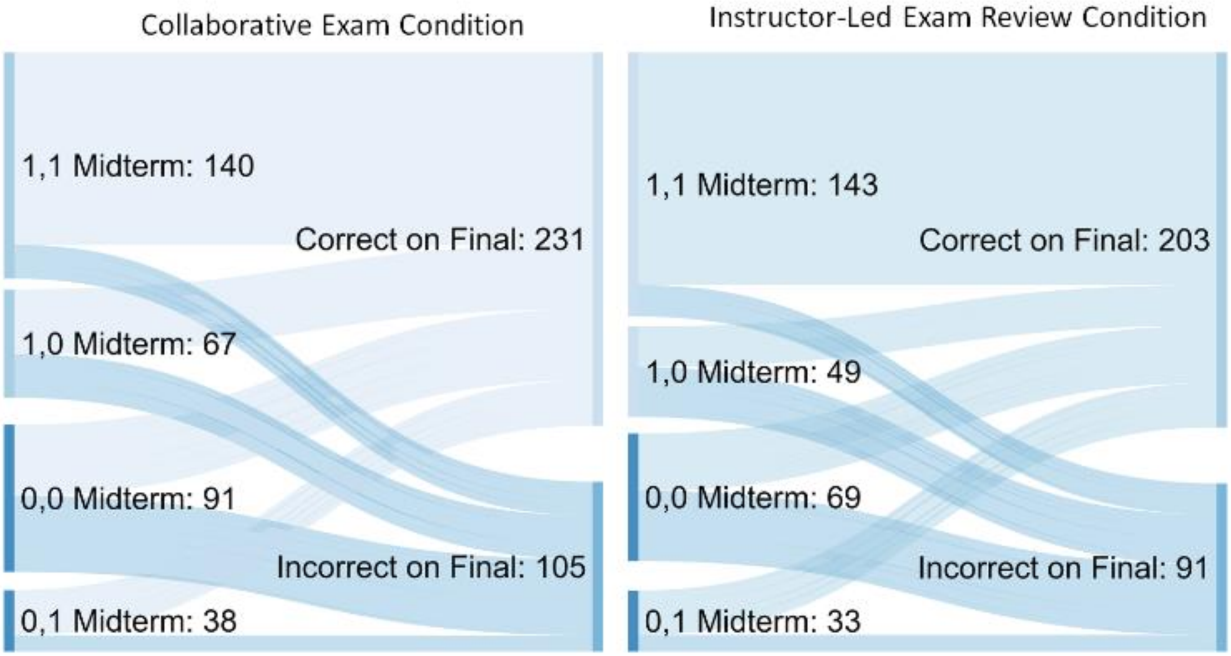


Figure 3.3. Sankey diagram showing shifts from midterm responses to final responses (in aggregate for both treatments).

A student-level analysis confirmed the result above, that no significant difference in student populations was detectable before the treatments. Specifically, for individual midterm responses, the average numbers of correct responses to target questions (per student) were $< N_{Target,Collab}^{Midterm} >= 2.7$ and $< N_{Target,Review}^{Midterm} >= 2.9$ for the collaborative exam and instructor-led exam review conditions respectively, with nearly equal variance. This difference is not statistically significant (t-test, p>0.05).

3.3.1.2. Analysis of Student Performance Post-Treatment.

In both conditions, student performance on the five target questions included on the final exam improved significantly. The average number of correct responses per student increased from the numbers reported above to $< N_{Target,Collab}^{Final} > = < N_{Target,Review}^{Final} >= 3.5$. This increase is statistically significant (t-test, p<0.05). While these averages are equal, the effect size (Cohen's d) of the collaborative exam condition, $d_{Collab} = 0.64$, is larger than that in the exam review condition, $d_{Review} = 0.44$, primarily due to a smaller variance in the student performance on the final exam in the collaborative exam condition ($\sigma_{\text{Collab}}^2 = 0.84, \sigma_{\text{Review}}^2 = 1.47$). Still, we find the moderate effect of the collaborative exam ($d_{Collab} = 0.64$,) to be reassuring given that the students in this condition did not receive any formal feedback from a source of "authority" (e.g., the instructor or exam solutions). In addition, the smaller variance in performance on the final exam in that condition seems to suggest that the collaborative exam treatment may be more equitable.

The two vertical bars on the right-hand side of each Sankey diagram in Figure 3.3 illustrate course-level student performance on the final exam. In both conditions, the fractions of correct responses were the same (69%), which is consistent with the student-level analysis of the shifts in the average number of correct responses per student. However, the Sankey diagrams reveal an interesting (and remarkably consistent) pattern in the "flow" of student responses from midterms to the final which allows for additional insights into nuanced aspects of the shifts in student performance. Nearly all [1,1] codes assigned to the midterm performance are also linked to correct answers to the target questions on the final exam suggesting that in the presence of conceptual understanding (indicated by the correct performance on the screening questions) student correct responses to the target questions appear to be stable over time.

Approximately half of the incorrect responses to the target questions on midterms ([1,0]+[0,0]) switched to correct answers on the final exam independent of performance on the screening questions or the treatment condition. Multiple interpretations are possible. One may expect that students who demonstrated correct conceptual understanding on a screening question would be more likely to improve their reasoning on the target question and therefore would be more likely to reason correctly on the final exam. The absence of this dependence may suggest that the improved performance on the final exam is a result of memorization. We argue, however, that the tendency to memorize is not a major factor in the observed improvement due to the following. First, as stated above, the exam solutions were not posted for review at any point during the semester. Second, the most significant improvement in performance was observed on the question included on midterm 1 (a few months before the final) with the smallest improvement observed on one of the questions included on midterm 3 (two weeks before the final). Third, and most importantly, a question-level analysis revealed that students appear to improve more on those questions that yielded higher performance on a midterm (at least 60% correct). This finding suggests that some intuitive ideas that remain even after formal instruction could be further addressed during (or after) summative assessment by implementing either a collaborative group exam component as a part of the assessment process, or by following up with an instructor-led exam review. At the same time, other intuitive patterns of reasoning appear to be less responsive to the "quick" interventions examined here and may require more targeted classroom instruction that takes into account both student conceptual difficulties and tendencies to reason intuitively. Our explorative data analysis helped identify contexts in which incorrect reasoning patterns do not appear to be responsive to the treatments described here (i.e., dynamics of circular motion, and work and energy); however, further research is needed to generalize to other instructional conditions.

3.4. Conclusions

In this project, we probed the impacts of a collaborative group exam and an instructor-led exam review in addressing student reasoning difficulties that remain even after classroom instruction. Results suggest that both approaches led to comparable improvements in performance on questions that tend to elicit incorrect intuitively appealing responses even in the presence of correct conceptual understanding. Nevertheless, we advocate for the implementation of the collaborative group exam approach in courses in which student group work is established to be a norm. A synergy between the classroom instruction and this assessment technique may provide additional benefits to student learning not examined in this study (e.g., developing social networks of support) as well as the potential for more equitable improvements in student reasoning.

Further, we argue that the two approaches examined in this study may serve not only as instructional techniques but also as research tools for identifying those patterns of incorrect student reasoning that require priority during instruction. Indeed, our results revealed that improvements in student performance appear to be higher on questions that already yielded fairly satisfactory performance on midterms. This suggests that perhaps the corresponding classroom instruction was already effective in addressing conceptual and reasoning difficulties so that some students simply needed a gentle nudge provided by the examined interventions. At the same time, on those questions that yielded less satisfactory performance on midterms, the effects of the interventions were minimized, thus suggesting the need for more rigorous targeted instruction.

4. EXAMINING THE IMPACT OF INSTRUCTIONAL INTERVENTIONS DESIGNED THROUGH THE LENS OF DUAL PROCESS THEORY ON STUDENT REASONING INCONSISTENCIES[3]

4.1. Introduction

Research in physics education has contributed to improvements in the teaching and learning of physics, largely by serving as a guide for the development of curriculum of undergraduate physics courses. Many research-based curricula have been successfully designed, tested, refined, and adopted nationally and internationally [McDermott and Shaffer, 2002; Laws 2004; Goldberg, Robinson, and Otero,2012; Prather et al., 2012; Novack et al., 1998]. Findings from ongoing research on the effectiveness of such materials continue to reveal impressive learning gains [Sokoloff and Thornton 1997; Shaffer and McDermott, 1992], even at sites other than the developers' institutions [Sharma et al., 2012; Kryjevskaia, Boudreaux, and Heins]. This work examines how Dual-Process Theories of Reasoning and Decision-making (DPToR), can inform the development, testing, and analysis of a single instructional intervention. An important goal of this work is to illustrate specifically how a theoretical framework, DPToR, can guide experimental design, as well as curriculum design. This work aims to explain a phenomenon observed by many physics instructors: students often reason inconsistently across conceptually identical contexts. These reasoning inconsistencies have been documented in the

[3] This chapter is largely based on a published paper that was co-authored by Mila Kryjevskaia, Mackenzie Stetzer, Beth Lindsey, Alistair McInerny, Paula Heron, and Andrew Boudreaux [Kryjevskaia et al., 2020]. Alistair McInerny held primary responsibility for data processing and analysis. He drafted the statistical analysis section of the paper and interpreted the results.

literature, and while they may stem from a variety of factors [Heckler, 2011; Kryjevskaia, Stetzer, and Grosz 2014; Elby, 2000; Kryjevskaia and Stetzer, 2012] their frequency and resilience to instruction are cause for concern.

Prior research has developed and tested pairs of conceptually identical questions with different surface features that tend to elicit different responses from students. In practice, the first (screening) question in the sequence requires straightforward application of one or more key physics principles, while the second question is presented in a slightly different context and has been empirically shown to elicit strongly held, intuitively appealing, incorrect responses. While students tend to reason correctly on the first question, many seem to abandon this correct line of reasoning on the second (target) question. Students instead tend to either draw upon intuitive ideas or apply formal knowledge incorrectly in an attempt to justify an intuitively appealing answer. For many students, some reasoning phenomena appears to inhibit the correct use of the correct, previously demonstrated, knowledge and skills in a different context.

The Dual Process Theory of Reasoning can explain these reasoning inconsistencies. As described section 3.1.2.1, DPToR models human cognition as two interacting, but largely distinct, processes: a fast, automatic, subconscious process (process 1, sometimes referred to as the heuristic process); and a slow, analytical, effortful process (process 2, sometimes referred to as the analytic process) [Kahneman, 2011; Evans, 2006]. These theories argue that productive reasoning relies upon both relevant conceptual knowledge (sometimes referred to as mindware) and the ability to mediate intuitive thinking by reasoning more analytically (sometimes referred to as cognitive reflection) [Stanovich, 2009; Frederick, 2005].   This work investigates the specific relationship between these two concepts, attempting to understand their distinct contributions to successful reasoning.

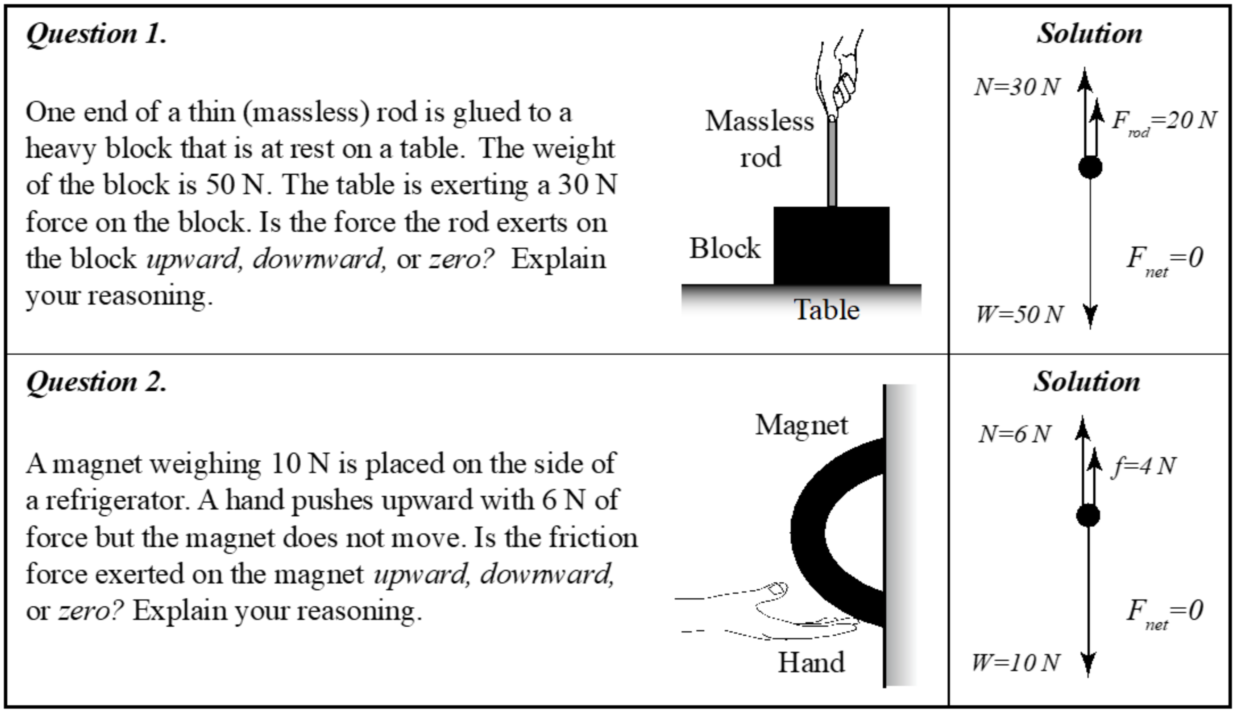
A three-stage instructional intervention was developed guided by DPToR and the associated constructs of mindware and cognitive reflection. This intervention was developed with three distinct goals in mind: 1) to disentangle the individual contributions of mindware and cognitive reflection skills, specifically with regards to productive reasoning on physics questions. 2) developing an intervention sequence that supports both the development of mindware and the engagement of cognitive reflection skills. And 3) assessing data to examine the overall improvement in student performance while also pinpointing the impact of the intervention on both mindware and cognitive reflection. This intervention was implemented in introductory calculus-based physics courses at three different institutions serving a broad spectrum of students and focuses on Newton’s second law.

The structure of the paper broadly follows the three overarching phases of our project. Section 4.2 provides motivation for the work and expand on elements of DPToR relevant to the chapter, such as mindware, and cognitive reflection. Sections 4.3, 4.4, and 4.5 describe the methodology, data collection, and analysis associated with Phases 1, 2, and 3 of the intervention, respectively. Section 4.6 discusses overall findings and broader implications for curriculum development. Finally, Section 4.7 summarizes these conclusions.

4.2. Motivation and Theoretical Framework

4.2.1. Inconsistencies in Student Reasoning

As previously mentioned, Target-Screening question pairs have been used to investigate inconsistencies in student reasoning. Figure 4.3 shows the specific question pair used in this study. In question 1, the block question, a thin, massless rod is glued to a heavy block at rest on a table. Students are told that the weight of the block is 50 N and that the table is exerting a 30 N force on the block. Students are asked to determine the direction of the force the rod exerts on the block, and to explain their reasoning. By applying Newton’s 2nd law, with F_{net}= 0 for the block, students can conclude that the rod exerts an upward force of 20 N. Students typically do well on this question with most students providing a correct answer and correct reasoning. (A more detailed discussion of results is found in section 4.3



Question 1.

One end of a thin (massless) rod is glued to a heavy block that is at rest on a table. The weight of the block is 50 N. The table is exerting a 30 N force on the block. Is the force the rod exerts on the block *upward, downward,* or *zero?* Explain your reasoning.

Question 2.

A magnet weighing 10 N is placed on the side of a refrigerator. A hand pushes upward with 6 N of force but the magnet does not move. Is the friction force exerted on the magnet *upward, downward,* or *zero?* Explain your reasoning.

Figure 4.1. Target-Screening question pair involving Newton’s laws.

Success on the block question demonstrates, in almost all cases, correct understanding of Newton’s 2nd law for an object in the at-rest. Therefore, this question serves as a screening question [Kryjevskaia, Stetzer, and Grosz, 2014]. Question 2, the magnet question, requires students to apply the same underlying reasoning in a situation with more complex surface features. These features frequently elicit intuitively appealing incorrect responses, establishing question 2 as a target question. In question 2, a magnet weighing 10 N remains at rest on the side of a refrigerator while a hand exerts an upward force of 6 N. Students are asked to determine the direction of the friction force exerted on the magnet. Physics experts recognize the magnet question as analogous to the block question; the magnet remains at rest, therefore the net force acting on it must be zero, and thus the friction force is upward. Students do not consistently approach this problem through the lens of Newtons laws, and performance on the magnet question is typically much poorer than performance on the box question.

Approximately one-third of students who answered the block (screening) question correctly answered the magnet (target) question correctly. Many students who correctly applied relevant concepts on the block question failed to apply the same reasoning on the target question. Most frequently, students incorrectly argued that the friction force acting on the magnet is downward because it must "oppose" the upward force exerted by the hand. In these responses students typically did not explicitly reference Newton's 2nd law or the gravitational force. As discussed further below, data suggest that the "friction-opposes-hand" argument is intuitively appealing and interferes with students' ability to access and apply the physics concepts used previously on the block question [Frederick 2005].

Similar reasoning inconsistencies have been reported in the literature [Heckler, 2011; Kryjevskaia, Stetzer, and Grosz, 2014; Kryjevskaia, 2019]. These results span a variety of conceptual domains in physics and occur both before and after research-based instruction. This work purports that these consistent and problematic reasoning inconsistencies are likely related to fundamental aspects of human reasoning and can be explained using DPToR as explained earlier.

4.2.2. Intuition and Reasoning

Despite its frequent use, the term "intuition" is complicated and somewhat undefined. Kahneman's parsimoniously defines intuition as "nothing more and nothing less than recognition" [Kahneman, 2011; Simon, 1992]. This operationalized definition allows us to apply the term intuition to both novice and experts while explaining the difference in outcomes between them. Under this definition, intuition can effectively be described as the first available mental model constructed and activated by a reasoner's process 1. As discussed previously, forming an initial mental model happens quickly, and generally below the level of conscious awareness. The formation and activation of a model occurs in response to contextual cues salient to the reasoner, although what elements are and are not salient are based on the reasoner's prior experiences. A physics expert and a physics novice, when presented with the same set of cues, are likely to construct different initial mental models, generally due to the differences in their relevant prior experience.

"Intuitive knowledge" has often been used by physics instructors and physics education researchers to refer to ideas that originate outside the physics curriculum. Along these lines, intuitive knowledge is sometimes framed as opposing formal knowledge (i.e., to the ideas presented in the physics classroom). Sabella and Cochran attribute poor performance on a mechanics task to either difficulties with the requisite formal ideas (i.e., Newton's laws) or to the use of an entirely different type of knowledge: "possibly a more intuitive response, or a p-prim [DiSessa, 1993] – the knowledge associated with Newton's Laws is not brought to the task" [Sabella and Cochran, 2004]. Similarly, Lising and Elby contrast "formal, classroom-taught reasoning" with "'everyday' and intuitive informal reasoning" [Lising and Elby, 2005]. This explains an important distinction between novices and experts; experts generally default to the relevant formal physics knowledge as their first available response, generally through repeated practice over many exposures. For an expert "intuitive knowledge" often is "formal knowledge", which is demonstrably not the case for a novice. A physics expert, with their extensive repertoire of prior experiences (including formal knowledge), is likely to recognize relevant cues when approaching a novel physics problem, and to then activate a productive mental model that is well-aligned with normative, formal physics knowledge. In contrast, a novice, with more limited experience, may key on unproductive or irrelevant cues, and activate a less productive mental model.

4.2.3. Dual Process Theory of Reasoning

Evans' heuristic-analytic theory [Evans 2006] (described in 3.2.1.1) provides a useful lens for understanding the phenomenon of inconsistent student reasoning. For example, imagine a novice student encountering a novel problem. As the student reads the question, the heuristic process generates an initial mental model. According to Evans' theory, only one mental model is considered at a time (singularity principle), and that model is selected based on its perceived relevance to the current task (relevance principle). The student's heuristic process focuses on contextual cues, which may or may not be relevant to a normative expert response. These cues strongly influence the mental model that is generated. If these cues are not relevant then they can be thought of as salient distracting features [Heckler 2011; Mamede et al., 2012; Osman and Stavy, 2006]. For example, on the magnet question the student might cue on "a push applied by a hand," activated by prior life experiences. This in turn may lead to the initial mental "friction resists an applied push."

Once the heuristic process has generated a mental model, the analytic process may intervene. Thompson argues that the default model is accompanied by a value judgment about its plausibility, known as the feeling of rightness [Thompson, Evans, and Cambell, 2009]. If the feeling of rightness is strong, e.g. the student is confident in their approach, analytic process intervention is unlikely. If this is the case then the analytic process is bypassed entirely and the student will respond with their first mental model [Thompson, Evans, and Cambell, 2013, Thompson, Turner, and Pennycook, 2011]. In the magnet question example, even though the student may be familiar with Newton’s 2nd law and is aware of the gravitational force, the feeling of rightness prevents any “red flags” from being raised (i.e., conflict detection). This direct path from first-available model to response is often described as cognitive miserliness, avoiding computationally expensive processing [Johnson-Laird, 2006; Toplak, West, and Stanovich, 2011].

Many questions posed in physics courses ask students to explain their reasoning, ideally necessitating a minimal level of analytic process intervention when articulating an explanation. The purpose of the analytic process is to ascertain whether or not the default mental model satisfies the task at hand (satisficing principle) [Evans, 2006]. The analytic process is inherently reflective with an aim of validating the default mental model. However, the effectiveness of any analytic engagement is impacted by the associated feeling of rightness. If a student has a strong feeling of rightness for a particular model, the engagement of the analytic process is likely to be superficial, even when an explanation is specifically asked for. Moreover, the analytic process is also subject to reasoning biases. Reasoners tend to be poor at searching for alternative mental models or generating counterarguments, and the analytic process is often driven by confirmation bias [Nickerson, 1998], shifting from analyzing to rationalizing. Even after the analytic process is engaged, an incorrect first-available mental model may still be the final response. However, if the analytic process is engaged, and a conflict with the first-available mental model is detected, the analytic process will hand the task off to the heuristic process for another mental mod, a new model is generated, and the reasoning cycle repeats.

As illustrated above, the nature of that very first available mental model plays a critical role in the student’s overall reasoning process. Background knowledge and intuition can be heavily impacted by the contextual cues, possibly preventing students from using relevant knowledge and skills, explaining how students can correctly apply concepts or skills in one context but fail to do so in a conceptually identical context.

4.2.4. Ingredients for Productive Thinking: Mindware and Cognitive Reflection

4.2.4.1. Mindware

Relevant knowledge is critical for successful reasoning. The term *mindware* is used in the cognitive science literature as an analogy to computer software, referring to the collection of "rules, knowledge, procedures, and strategies that a person can retrieve from memory in order to aid decision making and problem solving" [Stanovich, 2009]. Mindware contributes to the formation of the first-available mental model, as well as detecting reasoning red flags, and generating productive alternative mental models. The activation of mindware is not ubiquitous or necessarily consistent; similarly to software, a computer (brain) contains many more tools and much more information than are accessed on a given day. The context, framing, and even priming of a situation can affect which elements of mindware are brought to bear.

For example, physics experts will have many "modules" of physics-related mindware, such as Newton's 2nd law, ingrained much more deeply than a student learning Newton's laws for the first time. For the expert, the intuitive, first-available response to the magnet question is the same as the novice's formal reasoning. In addition to better trained intuition, experts are also more likely to detect and resolve errors in their reasoning, as they have the training and mental tools to test their mental models for those errors. Unfortunately, novices often lack the tools or the training, and as such can struggle to successfully assess their own mental models. Even if a student recognizes the conflict between their answer of "friction is downward" and Newton's 2nd law (perhaps because he or she is aware of the presence of a third force, due to gravity) while answering the magnet question, they may not have a robust mindware to shift from their original response to the accurate understanding that the friction force is upward. Sufficient mindware is necessary for successful reasoning.

4.2.4.2. Cognitive Reflection

Research has also revealed individual differences in the general tendency to critically evaluate one's mental models [Frederick, 2005, Tishman, Jay, Perkins, 1993]. Aside from differences in relevant mindware, some reasoners are more likely to evaluate the initial models put forward by their heuristic process. Mediating the heuristic process by reasoning more analytically is typically referred to as cognitive reflection. Even with comparable relevant mindware, the reasoner with a stronger disposition toward cognitive reflection is more likely to detect a conflict or error in their mental model. While tendency toward cognitive reflection does not ensure a reasoner will ultimately arrive at a correct response, it is more likely that the analytic process will function productively to identify models that are not satisfactory. On the opposite end of the spectrum, *cognitive miserliness* effectively denotes one's unlikeliness to cognitively reflect. Cognitive miserliness refers to the tendency of an individual to limit cognitive effort by relying on mental shortcuts or heuristics rather than engaging in more effortful, analytical thinking. The concept suggests that people often conserve mental resources by opting for quick, intuitive judgments instead of thorough and deliberate reasoning. This can lead to biases and errors in decision-making, as individuals may overlook relevant information or fail to consider alternative perspectives.

The Cognitive Reflection Test, developed by Frederick and validated and widely employed in psychology research, has been used to gauge this tendency toward cognitive reflection [Frederick, 2005]. Toplak, West, and Stanovich argue that the CRT is "a particularly potent measure of the tendency toward miserly processing," with a lower score indicating increased cognitive miserliness [Toplak, West, and Stanovich, 2011]. Many in the literature argue that the CRT is a measure of a reasoner's disposition toward "actively open-minded thinking," (i.e., the tendency to actively search for alternative answers) [Toplak, West, and Stanovich, 2011, Thompson, Evans, and Campbell, 2013; Pennycook et al., 2016; Campitelli and Gerrans, 2014; Stupple et al., 2013].

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

Figure 4.2. Three-item Cognitive Reflection Test. (Answers: 5 cents, 5 minutes, 47 days).

The original CRT, shown in Figure 4.2, consists of three questions. Each question generally elicits an intuitively appealing (but incorrect) response. Correct responses require that this intuitively appealing response is assessed by the analytic process, rejected, and then replaced with a correct answer. These questions (theoretically) require only basic mathematical and reading comprehension abilities and as such a correct answer should be dependent strictly on the respondent's tendency to evaluate their answer before accepting it and moving on. A single point on the CRT is awarded for each correct answer, and a score of 2 or 3 indicates a strong tendency towards cognitive reflection.

4.2.5. Summary

DPToR suggests that more developed mindware impacts physics performance in a variety of ways, but, when a student's first-available mental model is incorrect, cognitive reflection is extremely relevant when determining whether a correct final answer will be reached.

Questions developed through physics education research, such as the magnet question, often elicit a strong intuitive response, and thus frequently require students to successfully detect errors and override those errors to answer correctly. In this investigation, an intervention was developed and tested that was designed to enhance student performance on forces questions that elicit strong intuitive responses in order to better understand the roles of both mindware and cognitive reflection on student reasoning on performance.

4.3. Disentangling Mindware and Cognitive Reflection

4.3.1. Methodology

This section presents an overview of the context of the investigation and describes the project's overarching data collection and analysis methods. Methodological approaches unique to Phases 2 and 3 are discussed in the relevant subsequent sections.

4.3.1.1. Research Setting and Student Population

Data was collected in introductory calculus-based mechanics courses at three different universities in the US. These institutions serve a diverse range of students as measured by incoming Math SAT scores (see Figure 4.3). Students consisted primarily of physical sciences or engineering majors. At University A, the intervention was administered as part of the laboratory component of the mechanics course required by some, but not all, majors; data from this institution was collected over the course of two subsequent academic terms taught by the same instructor. At University B, the intervention was administered as part of a required laboratory component of the mechanics course. Multiple lecture sections, taught by different instructors, fed into this laboratory component. At University C, the intervention was administered to three lecture sections, taught by three different instructors in the same academic term. In two of those sections, administration of the intervention was identical, data from those sections is combined and referred to as “Section 1”. In the remaining section, “Section 2”, intervention stages 2 and 3 were not administered. In all courses, instructors used research-based, active learning strategies to various degrees.

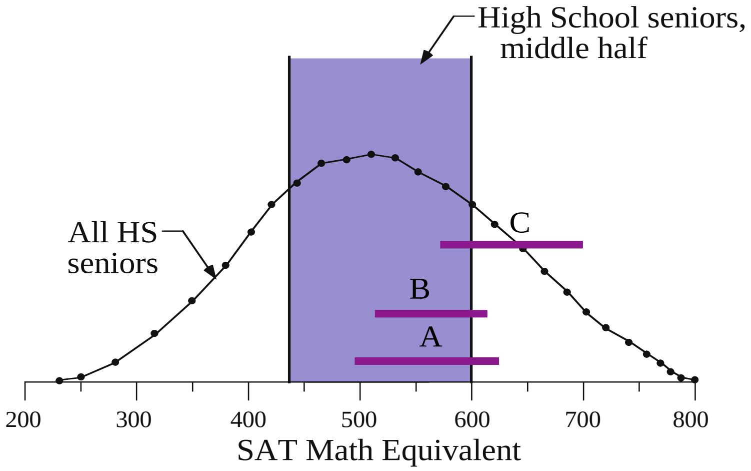


Figure 4.3. 25th -75th percentile of incoming math SAT scores compared to national values.

4.3.1.2. Data Collection

Data was collected using physics content questions related to the block-magnet sequence shown in Figure 4.1, across a variety of timing constraints: 1) data collected before any intervention had taken place (“pretest” data); 2) data that were collected as part of the intervention (“intervention” data), which included three distinct stages described in detail in section 4.4; and 3) data collected as part of a course exam (“post-test” data). 3-item CRT scores were collected for each population although at different times depending on the section as explained in Table 4.2. Although the same data were collected at each university, the intervention was administered differently depending on institutional constraints. Considering data from multiple instances of implementation mimics the effects of replication, producing more generalizable claims. These differences are described below and are summarized in Table 4.2.

Table 4.1. Data streams used in this study.

Data Stream:		University A	University B	University C (section 1)	University C (section 2)
Pre-test & Intervention Stage 1	Format	Online	Online	Online	Online
	Timing in relation to instruction	After all relevant instruction	Varied by section	Before all instruction on Forces and Newton's laws	Before all instruction on Forces and Newton's laws
Intervention Stages 2 & 3		In lab; answers submitted in groups via web form	In lab; answers submitted in groups via web form	In lecture; answers submitted individually using clickers	Not implemented
Post-test		Free-response on midterm exam	Free-response on final exam	Multiple-choice on midterm exam	Multiple-choice on midterm exam
CRT		Online near beginning of academic term	On paper near beginning of academic term	Online near end of academic term	Online near end of academic term

4.3.1.2.1. Pretest

The pretest was administered at all three institutions as part of a regular ungraded assignment in advance of a weekly lab or recitation. In each instance, the assignment was administered online outside of class, students were asked to complete the assignment individually, and credit was given based on effort rather than correctness of responses. For the pretest, students were shown the block-magnet question pair on a single page of the online assignment. Students were asked to answer each question in a multiple-choice format and used text boxes to explain their reasoning.

4.3.1.2.2. Cognitive Reflection Test (CRT)

At each institution, students received participation credit for completing the CRT, regardless of correctness. At University A, the CRT was administered at the end of a regular weekly online assignment early in the term. At University B, the CRT was administered on paper in the first week of the course. At University C, the CRT was administered at the beginning of a regular weekly online assignment near the end of the term. These differences are unlikely to affect the data, as student performance on the CRT appears to be relatively stable [Bialek and

Pennycook, 2018]. There is no evidence that CRT performance is impacted by instruction in a single physics course.

4.3.1.2.3. Data Analysis

For all data from streams 1-3, responses were coded as either “correct” or “incorrect.” A response was coded as correct if it contained the correct answer as well as associated correct reasoning.

Binary logistic regression was used to probe the relationships among variables of interest. Logistic regression analysis creates models that predict the probability of a particular dichotomous outcome (whether a student will respond correctly or incorrectly to a question) based on the value of various predictors (student’s score on the CRT or performance on a pretest question). The predictor variables can be categorical or continuous. The logistic regression algorithm fits a multiple linear regression algorithm for the log of the odds of an event:

$$\log(odds) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{1}$$

where the x_k are the predictor variables and the $\beta_k$are regression coefficients estimated by the algorithm. The odds of an event are given by:

$$odds = \frac{p}{1-p} \tag{2}$$

It can be shown from Eqs. (1) and (2) that the probability P(Y) of event Y occurring is then given by:

$$P(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_k x_k)}} \tag{3}$$

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

The assumptions of binary logistic regression are less stringent than for linear regression but include assumptions of (1) linearity, (2) independence of errors, and (3) a lack of multicollinearity. 1) For any continuous predictor, there must be a linear relationship between this predictor and the log(odds) of the outcome variable. 2) Separate cases of data should not be related, no single individual should appear multiple times in the data set (e.g., at different points in time). (3) Predictors should not be too highly correlated with one another. In this analysis, all assumptions were met satisfactorily.

4.3.2. Pre-Intervention Assessment: Data, Analysis, and Results

4.3.2.1. Student Reasoning Inconsistencies

The results of the analysis of student responses to the block and magnet questions administered before the intervention are included in Table 4.3. “Pretest” data was collected before relevant instruction. The percentage of correct responses on the block question ranges from 45% to 64%, as shown in Table 4.2. The overwhelming majority of students who answered the block question incorrectly did not respond using Newton’s 2nd law. The most prevalent incorrect response was cued by the notion that a massless rod cannot exert a force. Explained by one student; “the mass of the rod is zero, F=ma, so the force by the rod must also be zero.” A smaller fraction of students used the figure to argue that “the hand appears to push on the rod, so the rod pushes down on the block.” There is no evidence that these students attempted to answer the block question by applying appropriate formal physics knowledge.

Student performance on the magnet question was much weaker than on the block question. Across all student populations and instructional circumstances, only ~20% of students (or fewer) answered the magnet question correctly. Most students who demonstrated necessary mindware on the block question did not apply that same mindware to the magnet question. Less than 35% of students who answered the block question correctly also applied Newton’s 2nd law consistently on the magnet question. As discussed in section 4.2.1, the overwhelming majority of students who answered the magnet question incorrectly argued that “the force of friction must point down because friction opposes the applied force by the hand.”

Table 4.2. Average CRT score and Pre-intervention performance.

Student population	University A	University B	University C	University C
			(section 1)	(section 2)
	N=91	N=125	N=230	N=104
<CRT>	1.99	2.04	1.92	1.97
Performance on the block question (pre-intervention)	64%	45%	54%	59%
Performance on the magnet question (pre-intervention)	20%	9%	21%	21%
Performance on the magnet question of **only** those students who answered the block question correctly	N=58	N=56	N=124	N=61
	29%	18%	34%	33%

4.3.2.2. Interpretation of Inconsistencies Through DPToR

These results indicate the inconsistent reasoning tendencies discussed earlier: on two questions that both only require application of Newton’s 2nd law, student performance differs drastically. Students do not apply the necessary mindware in a consistent manner. Stanovich posits that reasoners reason incorrectly because a) they do not possess necessary mindware or b) they spend the least amount of time possible carefully processing information (i.e., not engaging in cognitive reflection). Disentangling mindware from cognitive reflection skills and investigating the relationships between both factors and student performance can help inform the use of instructional strategies. When designing instruction, it is important to ask: are students lacking in mindware, or cognitive reflection?

4.3.2.2.1. Does the Block Question Measure Mindware?

Stanovich defines mindware as the knowledge of rules and concepts required to reason and answer correctly [Stanovich, 2009]. The mindware necessary to solve the block and magnet problem are as follows: 1) F_{net} = ma, where F_{net} is the vector sum of all forces, and 2) for an object at rest (in which case a = 0), F_{net} must be zero. It is, however, challenging to design an instrument (or a task) that determines the exact degree to which a student possesses this mindware. Several approaches are worth considering. A simple approach would ask students to articulate Newton's 2nd law and the conditions necessary for an object to remain at rest. This approach, however, has multiple limitations. First, a student could possess declarative knowledge of Newton's 2nd law and may even memorize the conditions for an object to remain at rest. However, declarative knowledge and memorization are not sufficient, as students may not be capable of recognizing the applicability of this knowledge. Even basic scenarios involving more than two forces may not trigger the relevant mindware; additionally, students may not be able to identify all forces acting on the object or may struggle to execute vector addition. These possibilities suggest that asking students to consider a basic scenario, which requires the application of the identified knowledge, would be a more informative measure of student mindware. Those students who correctly analyze the scenario are regarded as having the necessary mindware. The block question above seems to fit these criteria; it presents a basic situation that involves more than two forces and requires the application of the same set of steps as the magnet question in order to arrive at an answer.

As noted earlier, the data suggest that some student responses to the block question are cued by the question's surface features (e.g., "a massless rod cannot exert a force" and "a hand touching the rod while the block is still in contact with the surface suggests that the hand is pushing on the rod"). Thus, it is possible that those students who do possess necessary mindware may be distracted from applying it. Examination of student responses suggests that only ~4% of the students who answered the block question incorrectly (10 students out of 251) provided a correct response with correct reasoning to the magnet questions, while ~70% of the students who answered the block question correctly gave an incorrect response to the magnet question (210 out of 299). These results suggest that it is highly unlikely that students will arrive at a correct answer to the magnet question if they failed to do so on the block question. As such, it is reasonable to treat the block as an adequate measure of whether or not a student possesses the mindware necessary to answer the magnet question correctly. At the same time, it is important to note that this measure of mindware does not indicate its strength (or robustness).

4.3.2.2.2. Identifying Relationships Among Mindware, Cognitive Reflection Skills, and Student Performance on the Pre-Intervention Magnet Question

Mindware alone does not predict student performance on the magnet question, which begs the question: what role does cognitive reflection play? Specifically, it is logical to anticipate that students who possess strong cognitive reflection skills are more likely to assess their reasoning recognize red flags and are therefore more likely to answer the magnet question correctly. Logistic regression models were generated for each student population to examine the relationships of both mindware (block question performance) and cognitive reflection (CRT score).

Table 4.3. Logistic regression models linking performance on the block question (mindware) to performance on the magnet question.

		University A			University B		
	Predictors	Coef.	Sig.	Exp(β)	Coef.	Sig.	Exp(β)
Performance on the block question pre-intervention	Intercept	-3.47	.001	.03	-4.22	<.001	.02
	Block Question	2.59	.014	13.3	2.69	.012	14.78
		University C (section 1)			University C (section 2)		
	Predictors	Coef.	Sig.	Exp(β)	Coef.	Sig.	Exp(β)
Performance on the block question pre-intervention	Intercept	-2.81	<.001	.06	-3.02	<0.001	.05
	Block Question	2.14	<.001	8.54	2.30	0.003	10.0

As discussed earlier, a correct answer on the block question appears to be necessary, but not sufficient, for a correct answer on the magnet question, which is confirmed by the results of the logistic regression models, as shown in Table 4.3. (Model statistics for universities A, B, C (sections 1 and 2), respectively: χ^2=11.4, p=0.001; χ^2=11.5, p=0.001; χ^2=30.7, p<0.001; χ^2=14.0, p<0.001). CRT score, however, does not appear to predict success on the magnet question, prior to the intervention.

DPToR provides a useful insight into the relationship of both mindware and cognitive reflection here. Analysis of students' responses shows that students frequently reasoned using "the force of friction opposes the applied force." This mental model is likely readily available (and therefore intuitively appealing) to students. This is likely due to how often it is correctly and appropriately applied while learning in introductory mechanics courses. For example, a block at rest on a horizontal surface pushed by a single horizontal force does experience friction in the direction opposite to the applied force. In fact, the concepts of static and kinetic forces of friction are commonly introduced in the context of a block on a horizontal surface experiencing a single horizontal push or a pull by an applied force [Dubson et al., 2006; Tipler and Mosca, 2007; Mazur, 2014]. Additionally, the notion that "friction must be overcome for an object to move" is consistent with everyday experience and therefore makes "intuitive" sense as well. Given that the heuristic "friction opposes the applied force" is useful both formally and intuitively is should not be surprising that many students immediately and confidently accepted the mental model that "friction opposes the applied force by the hand." Reinforced as it is, any engagement of the analytic process is likely to seem unnecessary. The strong "feeling of rightness" associated with this response likely deterred students from activating process 2 and recognizing the red flag; that this (intuitively appealing) mental model does not take into account the gravitational force acting on the magnet.

While the lack of relation between CRT score and magnet question responses seems at odds with DPToR, this is not the case. Student performance data on the magnet question suggest that many students, regardless of their CRT score, were confident in their incorrect but intuitively appealing responses, and were likely in a state of "cognitive ease" [Kahneman, 2011]: they have applied this model of friction successfully in many other situations and therefore were unlikely to perceive the need to question the applicability of that model. In other words, students were not likely to reconsider their answers or look for alternatives because their first available mental model was perceived as highly relevant and plausible. For those students who had not yet developed strong enough Newton's 2nd law mindware, the alternative mental model (i.e., friction opposing the applied force by the hand) remains most relevant and highly plausible and was therefore strongly perceived to be correct.

The lack of dependence on the CRT score suggests that those students who answered the magnet question correctly did so because their first available mental models were based on a correct and formal application of Newton’s 2nd law. We suspect that the mindware of these students was strong enough that they immediately and subconsciously recognized the magnet question as one that is solved by applying Newton’s 2nd law.

4.4. Intervention

4.4.1. Intervention Design: Rationale, Structure, and Methodology

The instructional goal of this intervention is to improve student performance on questions similar to the magnet question – namely, questions on which many students struggle to apply the relevant mindware that they have shown themselves to possess. The underlying focus was to use DPToR and the associated cognitive constructs of mindware and cognitive reflection as a framework while doing so. This section reviews this overall approach and describes the stages of the intervention, the rationale behind each stage, details of implementation and data collection at each site. Due to the need to incorporate the intervention into instruction at three different sites, all with different constraints, there were some necessary and important differences in implementation and data collection; though, again, these pragmatic differences in implementation strengthen the generalizability of these findings.

4.4.1.1. Overall Rationale: Mindware or Cognitive Reflection

As demonstrated by students’ pretest performance at all three institutions, many students (~50%) already possessed the mindware necessary to analyze the magnet question correctly; however, the vast majority of these students did not apply that mindware, likely due to the presence of an alternative, highly available mental model (“friction opposes an applied force”). This suggests a need to design instruction which helps students recognize the applicability of the existing relevant mindware to new situations, as well as to resolve inconsistencies between student responses on the two tasks (the block question and the magnet question). Drawing on DPToR produces two promising instructional approaches.

One approach could focus exclusively on improving student mindware, to the point of automatic and immediate recognition of the relevance of Newton's 2nd law to the magnet task [Stanovich, 2018]. This could be achieved by training students to apply Newton's 2nd law in a variety of situations in which an object remains at rest. After such training Newton's 2nd law would likely become a highly available mental model, leading most students to automatically apply to situations like the magnet question. In this case, the high availability of this mental model would overshadow all other alternative mental models and would render the engagement of cognitive reflection skills unnecessary for most students. This approach would hopefully solve the issue of student reasoning inconsistencies in the context of Newton's 2nd law but does not help students in other contexts. Cognitive reflection skills would not be expected to strongly impact student performance if students mindware is strengthened such that the first mental model agrees with normative models.

Alternatively, instruction could focus on helping students engage cognitive reflection skills to allow for more productive or successful analytical processing. In this case, students would recognize red flags in their reasoning and, because they have demonstrated adequate mindware on the block problem, they will hopefully correct their incorrect intuitive approach. This approach would produce post-assessment results that depend on both students' mindware and cognitive reflection skills. This approach is unlikely to help students who struggled with the block question, but should help resolve student reasoning inconsistencies, at least in this context. According to Stanovich, in the absence of relevant mindware, both the detection of red flags associated with a first-available mental model and productive exploration of alternatives are extremely unlikely [Stanovich, 2009; Stanovich, 2018]. Previous research has found that an intervention designed to support productive analytic engagement had little to no impact on students who lacked the requisite mindware [Spiers et al., 2019].

Arguments can be made for each of the two approaches above depending on the context and desired outcomes. For example, the mindware-focused approach may be favored in cases in which automatic application of mindware (e.g., algebraic or vector operations, unit conversion, or application of the right-hand rule) is critical for successful reasoning and problem solving [Mikula and Heckler, 2017]. However, it can be argued that it would be more pragmatic to focus on designing instructional interventions which focus on improving student reasoning via cognitive reflection will have context independent benefits. Cognitive reflection is a critical element of productive reasoning and is subsequently a valuable instructional goal. As such, this study endeavored to design an intervention that works to improve both student mindware and cognitive reflection by helping students: 1) develop or strengthen already existing mindware, 2) make connections between a situation in which students are fairly successful at applying mindware (the block question) and a situation that presents considerably more challenges (the magnet question), and 3) specifically activate and engage their cognitive reflection skills.

The intervention was structured in three stages. In Stage 1, students work individually on a task designed to raise awareness of similarities between the block and magnet questions. In Stage 2, students work in groups to analyze the original block and magnet question pair. In Stage 3, students work in groups through a scaffolded activity, answering questions designed to guide them to apply Newton’s 2nd law and to refine their ideas surrounding the alternative model “friction opposes the hand.” The complete intervention sequence was implemented at

Universities A, B, and C, although Section 2 from University C was only given the stage 1 intervention. For each university, the timing of the pretest, intervention, and post-test relative to instruction on forces is shown in Figure 4.4. Below, we describe the rationale for and implementation of each stage in detail.

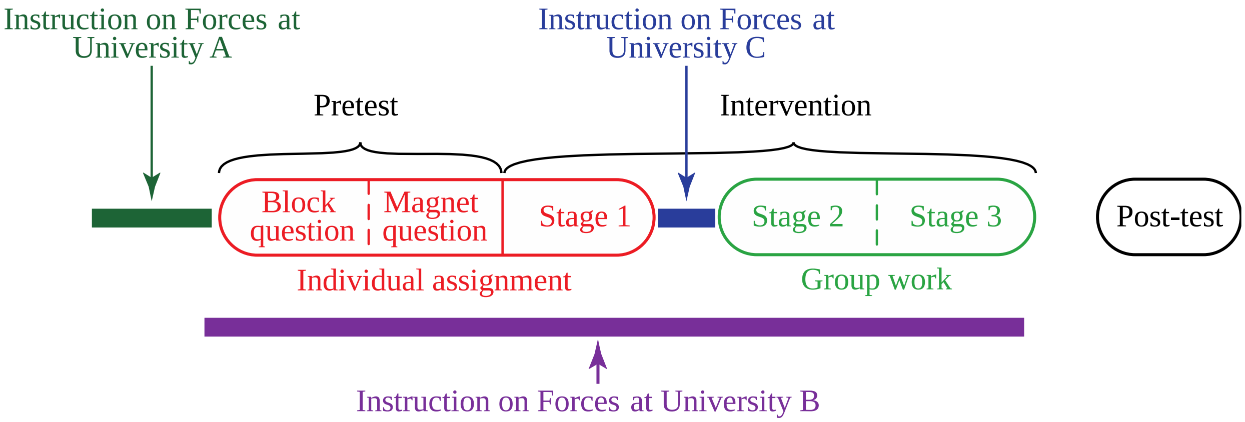

Instruction on Forces at University A
Instruction on Forces at University C
Pretest
Intervention
Block question
Magnet question
Stage 1
Stage 2
Stage 3
Post-test
Individual assignment
Group work
Instruction on Forces at University B

Figure 4.4. Intervention relative to instruction. Refer to Table 3.2. Students in "section 2" at University C did not experience stage 2 or 3 of the intervention.

4.4.1.2. Stage 1: Individual Work Intervention Designed to Raise Awareness of Similarities Between Block and Magnet Questions.

Stage 1 was implemented immediately after students completed the block-magnet question pair as part of the online pre-intervention pretest, which was administered outside of class in advance of a weekly lab or recitation. Students were taken to a new page in which they were provided with a (correct) solution to the block question, shown in Figure 4.5, and were asked to indicate whether or not they agree with the given solution and to explain their reasoning.

If an object remains at rest, its acceleration is zero; therefore, the net force acting on that object is zero. In this case, we know that two forces of given magnitudes act on the object: a downward force of magnitude 50 N and an upward force of magnitude 30 N, as shown at right. In order for the net force to be zero, a third force of magnitude 20 N must be directed upward.

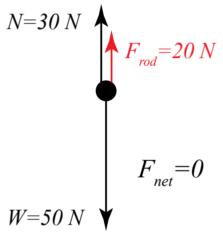


Figure 4.5. Feature-free solution to the block question used in Stage 1

The text of the solution was written using generalized terms that didn't explicitly reference the physical context, (e.g., the block, the specific forces acting on the block) to foreground the correct approach to analyzing forces acting on an object at rest. (The accompanying free-body diagram, however, did reference the specific forces acting on the block, as shown in Figure 4.5.) This feature-free solution does not directly connect the block and magnet questions, nor does it explicitly state that a similar analysis should be used in both cases. However, it was intended to gently nudge students, potentially triggering more productive engagement of the analytic process by raising awareness of similarities between the two scenarios. According to DPToR, if students recognize similarities between both scenarios while noting the differences in their reasoning approaches, it is more likely that potential red flags will be detected (e.g., by lowering their feeling of rightness and fostering a deeper engagement of the analytic process). This stage of the intervention was designed with the potential to help students mediate their initial, intuitive, thinking by helping them reason more analytically. It was also anticipated that exposure to a generic strategy for analyzing forces on objects at rest, and the opportunity to revisit the magnet question, would strengthen student mindware. To examine the effectiveness of the stage 1 intervention, students were asked whether they still agreed with their original answer to the magnet question, and to explain why they agreed or disagreed. If a student responded that they disagreed with their original response, they were given another opportunity to respond to the magnet question, and this response was recorded as that student's stage 1 response.

4.4.1.3. Stage 2: Group Work Intervention Involving the Block and Magnet Pair

Students worked in small groups in stages 2 and 3 of the intervention sequence. The group work format was expressly chosen as it allows for socially mediated metacognition [Kryjevskaia, Stetzer, and Le, 2014; Goos, Gilbraith, and Renshaw, 2002], in which group members help shape and guide each other's' thinking and reasoning, (as mentioned in section 3.1.2.2). The group, which is the effective unit of analysis, necessarily engages in self-assessment and self-regulation and draws upon its collective metacognitive knowledge while working collaboratively. The metacognition of the group is effectively externalized; with members generating new ideas, assessing information and approaches, disclosing their own thinking, requesting feedback on their own thinking, and monitoring their partners' thinking [Goos, Gilbraith, and Renshaw, 2002]. Through this collaborative thinking process, it is more likely that red flags will be identified and that intuitive ideas will be mediated via analytical reasoning. As mentioned earlier collaborative thinking has the potential to support the cognitive reflection of the group as well as opening the possibility that group work will help students learn the value of and strategies for cognitive reflection, which can in turn be employed while working individually.

This stage, which enabled students to revisit the block-magnet question pair in groups, was administered in person a short time after students completed the web-based stage 1. At Universities A and B, stage 2 was implemented in the associated laboratory component of the course. Students worked with their regular lab partners in groups of two to three students. Each group was tasked with discussing their approaches to the questions until a group consensus was reached, at which point a single group consensus response (including both an answer and reasoning) was submitted in web-based form for each of the two questions. As discussed earlier, this intervention was designed to foster cognitive reflection and promote consistency checking via socially mediated metacognition. It is also hoped that the opportunity to collaboratively revisit the block and magnet question pair could also help strengthen student mindware in some cases.

At University C, due to course constraints, the stage 2 intervention was implemented in lecture in section 1 as a series of multiple-choice clicker questions. Students received credit for a mixture of participation and correctness. As with Universities A and B, students had the opportunity to discuss each question with their neighbors, before voting on the answer. Because stage 2 was implemented through clicker questions implementation it was still possible for each student to respond individually rather than agreeing with their partner. In addition, only answers, not reasoning, were recorded. “Section 2” did not experience Stage 2.

4.4.1.4. Stage 3: Group Work Intervention Consisting of a Sequence of Guiding Questions

This stage of the intervention was originally designed exclusively for those lab groups that did not answer the magnet question correctly after stage 2. Using the same web-based online system, groups at Universities A and B who answered the magnet question incorrectly were automatically served a final sequence of questions intended to provide more scaffolding and step-by-step guidance to solve the magnet question. The stage 3 intervention was constructed to help students 1) recognize how Newton’s 2nd law can be applied to the magnet question, and 2) refine the intuitive notion that “friction opposes the applied force”. In the sequence of guiding questions, groups were shown the magnet question once again, but were asked to consider two different scenarios. In the first scenario, the hand is not exerting a force on the stationary magnet.

The second scenario is identical to the original magnet question, in which the hand is applying a force of 6 N upward on the magnet. Students were asked to draw a free-body diagram for the stationary magnet in both cases; they were then asked to determine the net force on the magnet in each case and to record their reasoning. Students were also asked what their responses suggest about the direction of the force of friction acting on the magnet in the first scenario and to explain their reasoning. In the absence of a hand, it was expected that students would conclude that friction does oppose the applied force, which, in this case, is the force of gravity. In this case, their intuition is consistent with the normative response; however, the force "opposed" by friction is gravity and not that from the hand. The inclusion of the second scenario, in which the force by the hand is added in the upward direction, allowed many students to refine their intuition by recognizing that "the force of friction must oppose the vector sum of the forces by gravity and the hand."

At University C, all students in section 1 participated in stage 3 in lecture as part of the clicker question sequence. Students discussed these questions and submitted their responses, which, ideally, reflected their groups' consensus responses. As with stage 2, only answers – not explanations – were recorded.

4.4.1.5. Intervention Data and Analysis Strategies

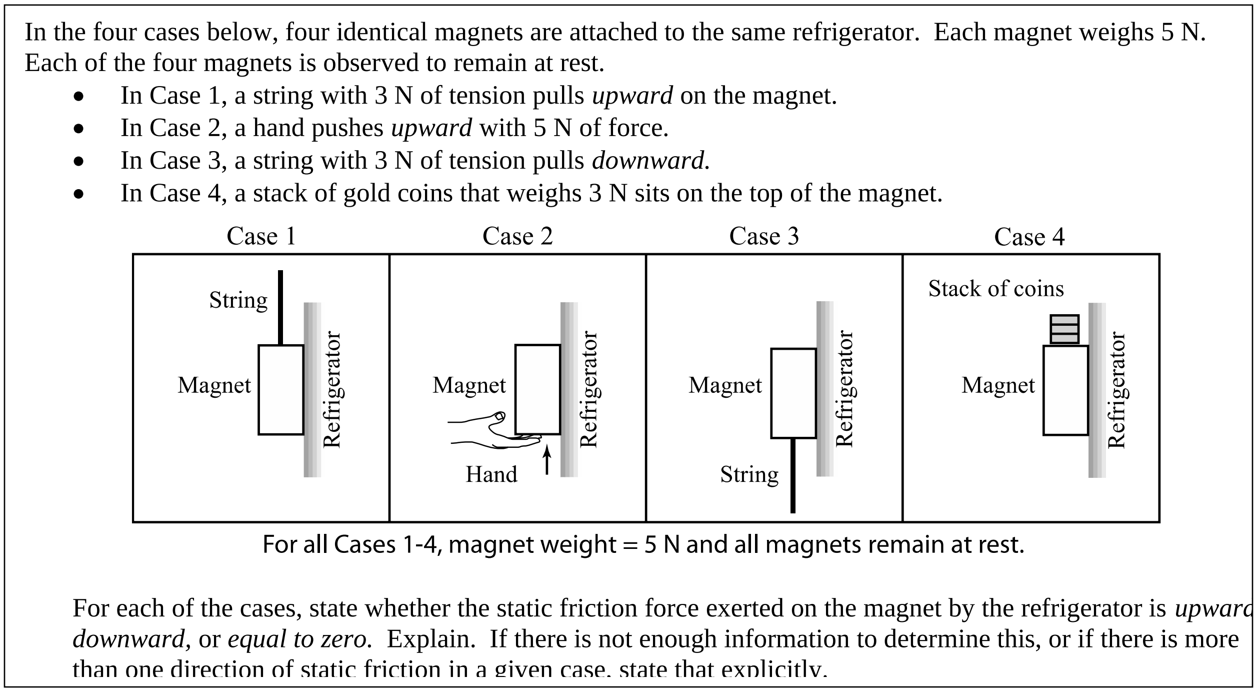
Student responses to the intervention questions were coded as either "correct" or "incorrect." At Universities A and B, a response was coded as correct if it contained the correct answer and was supported by the correct reasoning. For stages 2 and 3, a "correct" code was assigned to a student if that student was part of a group that had responded correctly, suggesting that the student agreed with the correct answer even though they did not necessarily generate that answer on their own. At University C, only stage 1 required students to explain their reasoning; responses from stages 2 and 3 were coded as correct solely based on correct answers.

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

4.5. Examining the Impact of the Intervention

4.5.1. Post-Intervention Data Collection

On the intervention post-test, a four-part question regarding a stationary magnet on a refrigerator was administered as part of a course exam at all three institutions (shown in Figure 4.6). Students were asked to identify the direction of the static friction force on the magnet in each of the four cases. The correct responses are that the friction force in Cases 1 – 4 are upward, zero, upward, and upward, respectively. At Universities A and C, this exam was a standard midterm exam administered to every lecture section. At University B, this exam was a final exam associated with the laboratory portion of the course, administered in the final laboratory period of the quarter. At Universities A and B, the post-test was administered in a free-response format, while at University C, the post-test was administered as a set of four multiple-choice questions.



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

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

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

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

Figure 4.6. Post test question administered to all students.

Responses were coded as “correct” if the student indicated the correct direction for the friction force in all four of the cases. At Universities A and B, such responses also had to be accompanied by correct reasoning. At University C, only the four correct answers were required. At Universities A and B, only about 2% of students gave four correct answers supported by incorrect reasoning, evidence that lack of reasoning (from students at University C) is not likely to be a cause for concern.

4.5.2. Mid- and Post-Intervention Assessments: Data Analysis and Results

As shown in Table 4.4, every intervention stage improved student performance by ~10%-30%. No single intervention stage is clearly more impactful than the rest. After the set of three interventions, most students gave a correct response to the magnet questions at least once at some point during instruction (or at least agreed with the correct answer); the success rate ranges from 74% at University A to 90% at University C (section 1).

Table 4.4 Student pretest, intervention, and post-test performance.

		University A	University B	University C (section 1)	University C (section 2)
Performance on the block question (pretest)		64%	45%	54%	59%
Performance on the magnet question	Pretest	20%	9%	21%	21%
	Stage 1	32%	20%	40%	40%
	Stage 2	49%	55%	59%	NA
	Stage 3	73%	81%	86%	NA
	Correct at any point during the intervention process	74%	85%	90%	40%
Correct on the post-test		62%	60%	77%	50%
Correct on the post-test of those who were correct at some point during instruction		69%	65%	78%	71%

The data shows a dramatic increase in student performance from pre- to post-test (from 9-21% correct on the magnet question to 60-77% correct on the post-test). However, it is important to note that most students in the study had at least three opportunities to consider the magnet question and modify their thinking. Students in "Section 2" at University C, who only completed stage 1, performed at the 50% level on the test compared to ~77% for the students at the same University who completed all three stages. It is striking that roughly 30% of students who, at some point, provided a correct answer to the magnet question reverted back to the incorrect notion that friction opposes the applied force. These results reinforce that this pattern of student reasoning is robust and persistent. The "quick fixes" of each interventional stage were relatively ineffective at individually correcting this trend. These results suggest that domain-general factors, such as cognitive reflection skills cannot be ignored, and that formal knowledge (or mindware) alone are necessary but not sufficient for correct reasoning.

Further analysis was conducted to answer specific research questions attempting to pinpoint whether the individual instructional interventions were functioning as intended , as well as why some students were successful on the post-test while others were not and. Specifically, did this set of instructional interventions 1) engage student cognitive reflection skills, 2) improve student mindware, and 3) to what extent did the intervention help students make connections between the block question (which most were able to complete successfully on the pretest) and the magnet question (which revealed reasoning difficulties)?

An estimate of post-intervention mindware was necessary to tackle the questions above. Prior to the intervention, student performance on the block question served as a suitable measure of mindware in relation to the magnet question. However, every intervention stage increased the rate of successful reasoning on the block question. As such, each intervention stage could (and likely did) change student mindware. Students' pre-intervention performance on the block question no longer adequately represented their post-intervention mindware. A parsimonious measure of students' post intervention mindware would be a correct response to the magnet question at any point. A correct response to the magnet question at any point indicates that a student either possesses the necessary mindware individually (during the pretest or stage 1) or was exposed to the correct reasoning during group work. Therefore, Any_point_magnet, mentioned earlier, indicates the presence or absence of the mindware necessary to answer the post-test correctly.

Student agreement with the correct answer to the magnet question during group work does not represent a student's ability to generate this response on their own. It is possible that a student was entirely disengaged from the discussion and its content. However, it is reasonable to argue that most students were conscious participants in the group discussions because this work was largely done in pairs. Even a student’s reluctant agreement with the correct reasoning suggests that the student was at least exposed to an alternative solution that contradicts their individual response, and that the student at least accepted the alternative solution as plausible.

Table 4.5. Logistic regression models predicting post-test performance as a function of mindware (measured by Any_point_magnet) and cognitive reflection (measured by CRT score). All models are statistically significant with p < 0.5. $\chi^2 = 12.12, df = 2, p < 0.01$, $\chi^2 = 17.4, df = 2, p < 0.01$, $\chi^2 = 9.4, df = 2, p < 0.01$, $\chi^2 = 14.5,\ df = 2, p < 0.01$ for University A, University B, University C section 1, and University C section 2 respectively.

Population	Logistic regression model
University A	$\mathrm{p} = \frac{1}{1 + e^{(1.7 - 1.4 * Any\_point\_magnet - 0.6 * CRT)}}$
University B	$\mathrm{p} = \frac{1}{1 + e^{(2.1 - 1.5 * Any\_point\_magnet - 0.6 * CRT)}}$
University C (Section 1)	$\mathrm{p} = \frac{1}{1 + e^{(0.3 - 0.9 * Any\_point\_magnet - 0.3 * CRT)}}$
University C (Section 2)	$\mathrm{p} = \frac{1}{1 + e^{(0.7 - 1.6 * Any\_point\_magnet)}}$

Table 4.5 contains logistic regression models that predict the probability of success on the post-test as a function of 1) mindware, as measured by the variable Any_point_magnet; and 2) cognitive reflection skills, as measured by CRT score. These models reveal that both mindware and cognitive reflection skills predict student performance on the post-test for students that completed all phases of the intervention. For students in University C (section 2), who only completed intervention stage 1, performance on the post-test is solely linked to mindware.

4.5.2.1. Why, Even After Targeted Intervention, Are Some Students Successful While Others are Not?

These results suggest that, after the intervention, students who possess both mindware and high cognitive reflection skills are much more likely to answer the post-test correctly compared to students who possess mindware but score zero on the CRT. For example, at University A, (given by$e^{0.6*3}$) a student who possesses the necessary mindware and receives a score of 3 on the CRT is 6 times more likely to answer the post-test correctly compared to a student who also possesses the mindware but scores zero on the CRT. These results are consistent with DPToR in that both factors, mindware and cognitive reflection, are important for successful reasoning (at least on those physics tasks that tend to elicit incorrect, but intuitively appealing responses).

4.5.2.2. Does the Intervention Engage Student Cognitive Reflection Skills?

As discussed in Section 4.3.2, student performance on the magnet pretest is a function of mindware only. After the intervention, student performance on the post-test is predicted by mindware and cognitive reflection skills for each population that completed all stages of the intervention. This result suggests that the intervention did engage student cognitive reflection skills. This is also evidenced by the fact that section 2, which did not experience stages 2 and 3 of the intervention, showed no dependency between post-test performance and CRT score.

4.5.2.3. Does the Intervention Improve Student Mindware?

Student performance significantly improved between the first attempt on the magnet question and the post-test. It is possible these shifts are entirely due to more productive engagement of students' cognitive reflection skills. However, comparing the number of students who successfully answered the magnet question on the first attempt to the number of students who provided correct answers to the magnet question at any point during the intervention provides a more accurate gauge of the cause of this change. Magnet question performance does not appear to correlate with CRT score. Therefore, improvement in performance is likely due to improvement in mindware. This argument does not suggest that cognitive reflection skills have no impact during the intervention, as discussed below. It does suggest, though, that the intervention appears to help improve mindware, independent of students' level of cognitive reflection skills.

4.5.2.4. Does the Intervention Help Students Make Connections Between the Block Question, and the Magnet Question?

Comparing student performance on their first attempt to answer the magnet question to their performance at stage 1 provides evidence that students make connections between their reasoning on the block question and magnet question. At University C, ~25% of the students who did not answer the magnet question correctly on the first attempt were able to reason successfully after stage 1. It is relevant to note that students who switched from incorrect to correct reasoning at stage 1, on average, had a higher CRT score compared to those who did not: $CRT_{switched}$=2.20, $CRT_{did\ not\ switch}$=1.85, Mann-Whitney U=5206.0, two-tailed p=0.019. This result supports the critical role of cognitive reflection skills in seeking reasoning coherence both during and after the intervention. At Universities A and B, the number of students switching from incorrect to correct responses at stage 1 was lower than that at University C, so it was not possible to verify the dependence of successful switching on CRT scores at Universities A and B.

4.6. Discussion and Implications

In this investigation, the data collection, analysis, and interpretation were all driven by the idea that two factors, mindware and cognitive reflection, impact student reasoning on physics questions. The design of the intervention sequence was informed by this idea, which is deeply rooted in DPToR. Indeed, the dual-process framework suggests that in order to switch from a first-available, highly compelling intuitive model to an alternative model based on formal knowledge, students must be able to 1) detect the conflict and 2) successfully mediate the intuitive thinking with analytical thinking that draws upon relevant mindware (i.e., engage in cognitive reflection) [Stanovich, 2009; Stanovich, 2018].

This study provides evidence that both factors play a critical role in student reasoning on a question that tends to elicit strong, intuitively appealing incorrect responses. The relationships among student performance, mindware, and cognitive reflection skills, however, are rather complex and nuanced.

4.6.1. Summary of Phase 1 Analysis

Students' pretest performance on the magnet question differed drastically from that on the block question. It is argued that the block question responses serve an adequate measure of the mindware necessary to answer the magnet question correctly. It is important to note that this measure does not indicate the strength of said mindware. Logistic regression models suggest that the presence of mindware is linked to performance on the magnet question, while cognitive reflection skills do not appear to predict success on that particular question. DPToR argues that the mindware of students who answered both questions correctly is likely to be quite strong, and it follows that that the magnet question cued a normative response as it's the first mental model. Without cognitive dissonance, the engagement of cognitive reflection skills was unnecessary.

Students who answered the block question correctly but failed to do so on the magnet question appear to possess the necessary mindware but fail to apply it to an analogous context. For these students, the first available mental model (i.e., friction opposes the force by the hand) appears to hold strong enough intuitive appeal that students did not perceive the need to consider alternatives. Cognitive reflection skills did not appear to be relevant or predictive of the outcome on the magnet pretest.

4.6.2. Summary of Phase 3 Analysis

Student performance on the post-test revealed significant improvements after the instructional intervention. A new measure of mindware was established in order to examine the impact of the intervention on reasoning and to probe the relationships among post-test performance, mindware, and cognitive reflection skills. Data suggests that the intervention did impact both mindware and cognitive reflection. Student performance improved meaningfully at each stage of the intervention. In addition, the intervention also helped at least some students build connections between the reasoning approaches required to answer the block and magnet questions correctly. Logistic regression models suggest that, after the intervention, student performance on the post-test is linked to both the presence of mindware and strength of cognitive reflection skills. Even after controlling for mindware, students with greater tendency to cognitively reflect were more likely to answer the post-test question correctly. These results are consistent with DPToR in that both factors, mindware and cognitive reflection, are important for successful reasoning.

4.6.3. Limitations

It is encouraging that these empirical results are consistent with theoretical predictions. Despite this there are two limitations/concerns associated with these findings. First, given that the post-assessment involved near transfer from the intervention itself, it is unclear whether the single intervention presented here will help students become attentive to conflict detection in other contexts. This issue is the focus of an ongoing investigation.

Second, while the theoretical framework used here assumes causality between cognitive reflection skills and student performance, the data presented here does not explicitly establish this causal link. It does logically follow that in the presence of necessary mindware, an improvement in cognitive reflection skills is likely to result in improvements in student performance. Current research on cognitive reflection skills suggests that student performance on the CRT is quite stable [Bialek and Pennycook, 2018] and it is unclear how to directly improve student cognitive reflection skills in general. Because of this, students who enter physics courses with relatively weak cognitive reflection skills will likely continue to perform poorly on questions that elicit strong, intuitively appealing responses compared to their peers who possess stronger cognitive reflection skills.

4.6.4. Insights into the Nature of Student Reasoning in Physics and Implications for Physics Instruction

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

Intuitive processing is deeply ingrained. Quick and isolated content specific fixes do not seem likely to completely resolve student reasoning difficulties in contexts that elicit strong, intuitively appealing responses. These analyses suggest that no single intervention stage was more impactful than the rest, and each stage produced a small but meaningful improvement in student performance. Even after the completion of the entire intervention sequence, one third of students reverted from the correct, normative response, to incorrect, intuitive reasoning on the post-test. This finding highlights the deeply ingrained nature of intuitive processing. Perhaps the most powerful aspects of DPToR are the assertions that 1) the first available mental model serves as an entry point into any reasoning path and 2) process 1 cannot be turned off. These fundamental aspects of reasoning highlight the importance of expertise either in the form of strong mindware (improving the frequency of correct first mental models), strong cognitive reflection (improving the frequency and quality of process 2 intervention), or the combination of the two.

4.6.4.1. Limitations of Instruction Exclusively Focused on Mindware

These results provide evidence that instruction with a primary focus on improving mindware might fail to help students overcome errors in their fundamental reasoning patterns. The inconsistency in student reasoning on the block-magnet pair prior to the intervention suggests that mindware alone is not enough to help students detect reasoning red flags or to successfully mediate intuitive responses. While domain-specific expertise is necessarily multi-faceted, the ability to detect and mediate intuition-based responses is an important ingredient for developing expertise. Designing curriculum with the sole purpose of improving mindware is likely to be less effective at helping students reason productively on many kinds of physics questions.

4.6.4.2. Need for Instructional Focus on Cognitive Reflection

These findings suggest that novices may not spontaneously engage in cognitive reflection on physics questions. On the pre-intervention magnet question, the lack of dependence between student performance and cognitive reflection skills (a similar result was observed on the post-test for those students who did not complete all stages of the intervention) suggests that many students were not able to recognize or act upon red flags in their reasoning, even in the presence of the requisite mindware. This suggests that interventions designed to help students recognize instances of intuitive thought and to help them develop productive approaches for mediating such thoughts are necessary for improved student reasoning, but not sufficient for completely resolving inconsistencies.

4.6.4.3. Impact of the Deeply Ingrained Nature of Intuitive Reasoning

The complex interplay between mindware and cognitive reflection skills may also explain why it is common for students who perform well in classroom activities (e.g., clicker questions, group work) to be less successful on nearly identical individual situations. The socially mediated metacognition that occurs during classroom activities allows students to process new information but may create a false sense of security regarding individual ability. The decline in student performance can be frustrating to instructors and students. According to Stanovich, unless mindware has “been practiced to automaticity” such that it “can automatically compete with (and often immediately defeat) any alternative non-normative response,” there is always room for error in applying that mindware for a student who has not yet developed the habits to recognize intuitive thought and/or to trust formal knowledge enough to override a strong intuitively appealing response. It may benefit students to remind them of the specific role of socially mediated metacognition during classroom activities, emphasizing the value of cognitive reflection in addition to any benefits of said activities.

Experienced instructors often hear frustrated students sharing that they studied hard and knew what concepts were relevant to each exam question, but that they frequently second guessed themselves (in the wrong direction). The phenomenon of second guessing, while frustrating to students, is a natural aspect of learning how to reason productively. Second guessing implies growth; a student who second guesses themselves has shown that they possess 1) the mindware to approach a problem multiple ways and 2) the cognitive reflection skills to have considered alternatives, rather than immediately accepting one's first mental model. While these students present some growth, they also have not yet developed the habits and skills mentioned above, which would also them to overcome their second guessing. It could benefit students to highlight the growth that second guessing represents, a difficult but important stage between being confidently incorrect and being confidently correct.

This work might lead one to wonder if instruction which makes the dual nature of human cognition explicit and visible to students would impact student learning of physics, possibly in ways that extend beyond improvements in student performance (see chapter 5). For example, highlighting that inaccurate first mental models are effectively a necessary element of novice reasoning could reduce the pressure students feel to "be right", and encourage students to feel comfortable examining their own responses and thinking as a natural part of learning. Feelings of student inadequacy are a common concern in research-based materials that elicit incorrect intuitive responses as a part of the instructional cycle. Explicit discussion that such incorrect responses stem from the dual nature of human thinking, and that the fast, survival based intuitive processing cannot be turned off may help alleviate such concerns. Developing an awareness of one's own thinking paths, as well as the ability to recognize red flags, are critical steps toward the development of expertise in any subject, and supporting students throughout the instructional process as they strengthen these metacognitive skills could provide significant benefits to students and teachers.

4.7. Conclusion

This study implemented and assessed the impact of a single intervention sequence at three different institutions, focused on the application of Newton's 2nd law to objects at rest, using DPToR to inform the design, assessment, and analysis efforts of the intervention. Dual-process theories were used to specifically design the intervention sequence with the aim of delineating and supporting two essential factors of productive reasoning: mindware and cognitive reflection. Data analysis indicates that the intervention improved students' mindware, engaged students' cognitive reflection skills. Specifically, these findings reinforce the benefit of both cognitive reflection and mindware, illustrating that cognitive reflection skills alone are insufficient. Despite this, it can be seen even relatively modest efforts to support cognitive reflection in tandem with the development of mindware can bolster student performance, helping students mediate intuitive thinking (which cannot be turned off) more productively and purposefully by reasoning more analytically. While these benefits are meaningful, interventions along these lines are unlikely to resolve the issue of reasoning inconsistency completely. This analysis has shown, consistent with the findings of other studies, that these kinds of intuitive processing errors occur despite strong content understanding. It is important to help students recognize that critically evaluating one's intuitive thoughts is integral to scientific thinking. In fact, intentional reflective thinking is often necessitated by the nature of human reasoning. Students should be encouraged to view errors as an inherent part of the thinking process, alleviating stress and encouraging reflection. Understanding human reasoning and using that understanding to design and assess teaching practices are important aspects of improving the way physics is taught and learned.

5. EXAMINING THE EFFECT OF EXPLICIT DUAL PROCESS THEORY OF REASONING INSTRUCTION ON STUDENT REASONING INCONSISTENCIES [4]

5.1. Introduction

In higher education institutions in the U.S., programs of study in science, technology, engineering, and mathematics (STEM) generally follow a sequential structure in which students study topics and complete courses in a prescribed order [Borda, Haskell, and Boudreaux; 2022]. Presumably, this structure arises from an assumption that students will draw on previous learning to develop understanding in new subject areas. This assumption applies at smaller scales, such as when students apply knowledge of Newton's laws learned in introductory mechanics to understand Coulomb interactions at the beginning of the subsequent E&M course, as well as at larger scales, such as when students apply energy concepts from introductory physics and chemistry to understand metabolism in an upper division biology course.

Research in physics education, however, has revealed a robust learning phenomenon: students often demonstrate conceptual understanding by reasoning successfully on one task but then seem to struggle on a related task. In other words, novice physics learners are often inconsistent in their reasoning, even when presented with conceptually similar tasks administered in proximity. Such reasoning inconsistencies could undermine the sequential structure of classroom teaching described above. Instructors might become frustrated by inconsistent student

[4] This chapter is largely based on a submitted paper that was co-authored by Alistair McInerny, Andrew Boudreaux, and Mila Kryjevskaia, [McInerny, Boudreaux, and Kryjevskaia, submitted]. Alistair McInerny held primary responsibility for data processing, analysis, and drafting of this text.

reasoning, wondering why their students do not seem to be making use of the concepts developed in the course or in previous courses. Students may also become discouraged when they make errors on tasks involving concepts they feel they have already learned. A major challenge for education researchers, curriculum developers, and instructors is to understand the nature and origin of reasoning inconsistencies and to find ways to help students resolve them. Indeed, consistent application of fundamental principles can be regarded as a hallmark of the transition from novice to expert.

In recent years, physics education researchers have drawn on Dual Process Theories of Reasoning (DPToR) to account for some of the student reasoning inconsistencies identified through empirical research [Heckler and Bogdan, 2018; Kryjevskaia, Heron, and Heckler, 2021; Speirs et al., 2019; Spiers et al., 2023; Lindsay, Stetzer, and Spiers, 2023; Lindsay, Nagel, and Savani, 2017]. According to DPToR, human cognition occurs through two processes [Evans, 2006; Kahneman, 2011]. Process 1 is fast and automatic and serves as the entry point for all reasoning. Upon encountering a novel challenge, such as a new physics problem, process 1 responds through simple recognition, zooming in on whatever elements are familiar, and quickly generating a provisional mental model for the situation. If this model is compelling, it will give rise to an answer or output without further review. Process 2 is slow and effortful and is only sometimes activated. If the provisional model is accompanied by doubts or red flags, then Process 2 may be engaged to evaluate the model. However, research has shown that even if process 2 is engaged, reasoning biases may prevent the detection and override of errors [Gilovich, Grifin, and Kahneman, 2002]. Thus, in some cases, a flawed mental model and accompanying non-normative output can persist even though Process 2 has come online to perform an evaluation. In other cases, a Process 2 analysis may lead to the rejection of the initial model, triggering the production of a new model and an iterative cycle of cognition.

A central feature of the dual process account of inconsistencies in students' physics reasoning is that a physics learner, even when giving an incorrect answer to a physics question, may well "know" relevant concepts and "possess" relevant reasoning abilities. This knowledge, however, may remain inactive [Heckler and Bogdan, 2018; Kryjevskaia, Stetzer, and Grosz, 2014; Heckler, 2011; Spiers et al., 2021]. Although available, it may be inaccessible to the learner when a compelling mental model generated through process 1 has been validated through a shallow engagement of process 2. DPToR suggests that helping learners slow down their cognition could support deeper engagement of process 2, including the activation of available, but previously latent resources to review a potentially flawed initial response.

Many instructional approaches could be devised to slow down students' thinking. Furthermore, the central role of recognition in the operation of process 1 suggests that different physics contexts could require different approaches. It could be argued that a sensible first attempt at designing physics instruction informed by DPToR is to simply tell students about the dual nature of human cognition, and how and why to slow down their thinking as they encounter novel problems in their physics course. If students are aware of the cognitive processes that can naturally lead to inconsistent and erroneous reasoning, they might actively mitigate the tendency to accept without scrutiny the first mental model that springs to mind. Knowledge of fast and slow cognitive processing might lead students to actively invest in the mental effort needed to sustain analytical evaluation of their first available ideas. Sustained process 2 evaluation can then provide opportunities for students to access concepts, principles, and other reasoning resources relevant to a given situation but not necessarily cued by the quick and automatic process 1.

As researchers who work with DPToR and are also active classroom physics instructors, we have found our students to be quite interested in scholarship on human cognition. Discussions of dual-process theories engage students, particularly the idea of a fast-thinking process that can "hijack" cognition. In addition, DPToR provides a natural explanation for the occurrence of mistakes during learning. Discussing DPToR with the students can help normalize such mistakes and, hopefully, reduce negative connotations often associated with making errors. Our research interest in dual-process theories, together with our positive experience discussing DPToR with our introductory physics students, led us to consider the possibility that explaining dual processing to students might "inoculate" them against some of the reasoning pitfalls associated with the interactions between the two processes and improve performance on physics questions known to elicit intuitive but incorrect responses.

This article reports on a controlled study to investigate the efficacy of incorporating discussions about the duality of human reasoning and its implications for physics learning into physics instruction. Our central goal was to probe to what extent, if at all, the incorporation of explicit discussions about the duality of human reasoning improves student performance on physics tasks that tend to elicit intuitively appealing but incorrect responses. We hypothesized that explicitly discussing the reasoning pitfalls associated with the quick and automatic process 1, as well as the cognitive biases of process 2, could boost students' performance on a variety of topics in physics. Moreover, if successful, such instruction could also have positive impacts on student learning and attitudes toward physics that extend beyond improved performance.

Section 5.2 presents dual-process theories in further detail, including their connections to reasoning in physics. Section 5.3 describes the setting for and design of the teaching experiment, including the classroom intervention itself. Findings are discussed in Section 5.4 and discuss the implications of those findings along with possible avenues of further research in Section 5.5.

5.2. Theoretical Framework and Prior Research

5.2.1. Dual-Process Theories of Reasoning

Dual-process theories of reasoning posit two modes of cognitive processing, which operate separately but do interact [Evans, 2006; Kahneman, 2011; Evans and Stanovich, 2013; Thompson, Evans, and Campbell, 2018; Gilovich, Griffin, and Kahneman, 2002]. Process 1 is fast and always active. It operates beneath the level of conscious awareness without a sense of mental effort. Process 1 continuously scans the environment, evaluates new situations for potential threats and opportunities, and prioritizes rapid evaluation that is at least somewhat accurate. According to DPToR, when faced with a novel challenge, a reasoner's response always begins by engaging Process 1. Salient contextual features cue an associative cascade of plausibly relevant ideas, explanations, and heuristics. This associative network is driven by whatever features of the situation the reasoner recognizes as familiar, and quickly gives rise to a first-available mental model of the situation. Process 1 is sometimes referred to as the intuitive or heuristic process. Importantly, Process 1 is subconscious, and many students may be unaware of how they develop their first mental models because they do so reflexively. Indeed, Herbert Simon has described intuition as "nothing more than recognition" [Simon, 1992] and that recognition is an important part of what separates novice and expert reasoners, both in terms of process 1 and process 2.

In any domain, experts tend to generate fairly accurate initial models, based on diagnostic cues, while novices generate relatively less accurate initial models, often based on spurious cues. For example, when asked to account for the constant speed motion of a book being pushed across a horizontal table, “ongoing motion” might be the most salient feature for a physics novice, leading to the activation of a mental model that “continuous motion requires a continuous push.” This mental model might then lead the novice to explain that a constant-strength, forward net force accounts for the book’s motion. A physics expert, however, might cue on “constant speed” as the most salient feature, leading to the activation of a mental model that “constant velocity means balanced forces.” This mental model might then lead the expert to explain that the forward push is canceled by a backward friction force, and that a net force of zero accounts for the constant speed of the book.

Relevant prior knowledge of concepts, procedures, and skills constitute what is referred to as “mindware” [Stanovich, 2009]. In the DPTOR framework, mindware, unlike computer software, is not regarded as either “installed” or “absent.” Rather, there can be different levels of instantiation of mindware. For experts, relevant mindware is deeply instantiated, making it highly accessible and easily activated when encountering a new challenge [Stanovich, 2018]. In contrast, a novice may “possess” relevant mindware, but with shallower instantiation. As a result, the mindware may be available, but not readily accessible in the moment. The relevant mindware may not be as strongly linked in the associative networks of novices, remaining dormant during the initial process 1 engagement. The level of instantiation of relevant mindware, and its corresponding accessibility helps explain differences in accuracy of the first available mental models generated by experts and novices, although other elements of reasoning also differentiate the two.

Reasoners often accept the first available mental model as correct without further scrutiny if that model is intuitively appealing and highly plausible. This progression, from the initial model directly to a final output, has been referred to as the path of cognitive frugality (see Figure

5.1a) [Kryjevskaia, Heron, and Stetzer, 2021]. To catch a mistake, the slow, effortful, and rule-based process 2 must intervene to disrupt the path of cognitive frugality and evaluate the output of process 1, as shown in Figure 5.1b.

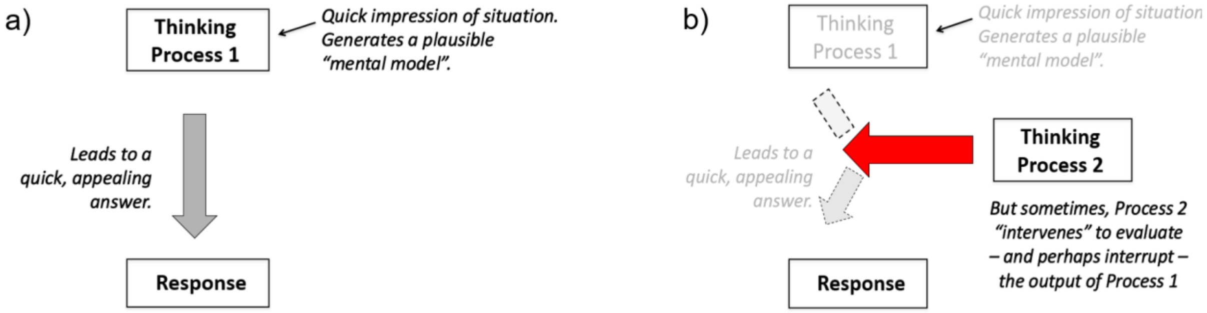

a)
Thinking Process 1
Quick impression of situation. Generates a plausible "mental model".
Leads to a quick, appealing answer.
Response
b)
Thinking Process 1
Quick impression of situation Generates a plausible "mental model".
Leads to a quick, appealing answer.
Thinking Process 2
But sometimes, Process 2 "intervenes" to evaluate – and perhaps interrupt – the output of Process 1
Response

Figure 5.1. Representation of reasoning pathways according to DPToR: a) Path of cognitive frugality and b) engagement of process 2.

The analytic thinking process, process 2, is subject to a variety of biases of its own. As such, the analytic process can come online to review a flawed initial model without successfully detecting any errors. For example, reasoners tend to prioritize evidence supporting their first available mental models, rather than searching actively for alternatives (i.e., confirmation bias) [Nickerson, 1998]. Furthermore, even if a reasoning red flag is detected, reasoners may be unable to override the first available mental model successfully if the relevant mindware is absent or weakly instantiated.   Even if a reasoner has significant doubts about their initial mental model, they may still find it difficult to generate a convincing alternative if their mindware is weak. Weakly instantiated mindware may also lead to "second-guessing" behavior, perhaps familiar to experienced instructors, in which students struggle to choose between an intuitive answer and a response based on formal knowledge learned in class that contradicts their intuition.

Cognitive reflection refers to the initial stage of type 2 engagement, in which a process-one-generated mental model comes under scrutiny. Some reasoners are more likely to detect mismatches between their initial mental model and the context at hand, while others may be more likely to accept their intuitive responses without further scrutiny. Research suggests that the tendency to recognize when to engage in process 2 analysis is linked to the reasoner's disposition toward cognitive reflection [Frederick 2005; Toplak, West, and Stanovich, 2014]. Figures 1(a) and 1(b), together, provide a simplified model of dual-process cognition. As discussed further in Section 5.4, these figures were presented to students as part of the teaching intervention.

5.2.2. Cognitive Reflection Test

We use the Cognitive Reflection Test (CRT) as a measure of the tendency to mediate intuitive thinking by reasoning more analytically. The CRT was developed specifically for that purpose, and is widely used in cognitive psychology [Frederick, 2005; Pennycook et al., 2016; Campitelli and Gerrans, 2014; Liberali et al., 2012; Stagnaro, Pennycook, and Rand, 2018; Meyer, Zhou, and Frederick, 2018]. This three-item instrument is designed to elicit intuitively appealing but incorrect responses that can be readily evaluated, rejected, and replaced with correct answers upon brief reflection. Solving CRT items requires only basic knowledge and skills (i.e., mindware such as arithmetic). For example, consider the first CRT item: "A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" For most reasoners, an intuitive answer of "10 cents" quickly comes to mind. Many accept this answer as correct without further consideration. However, those who reflect on their answer often realize – without needing additional insight into the relevant arithmetic – that the correct answer is actually "5 cents." We argue that students enrolled in college-level calculus-based physics courses possess the mindware necessary to answer the CRT questions correctly.

Therefore, in this study, we use CRT score as a measure of the tendency toward cognitive reflection. The original CRT consists of 3 questions, and due to the frequency of its use some experts are concerned that future respondents may already be familiar with these questions, invalidating their ability to test for reflectiveness. In order to avoid that possibility work has been done to develop new "CRT" questions, and 4 new questions have been developed. These questions can purportedly be used separately from the CRT or in conjunction with it. In this study the "CRT 7," the combination of the original CRT and the new one, has been used, in order to examine the viability of and differences between these methods of probing reflectiveness.

Prior studies in physics education research (PER) have identified relationships between CRT scores and student performance in physics. For example, Gette and Kryjevskaia designed and administered a sequence of screening-target questions related to Newton's third law [Gette and Kryjevskaia, 2019]. The screening questions involved applying Newton's third law to scenarios that do not tend to evoke strong, intuitive, incorrect responses. The target questions, however, required applying Newton's third law to a two-body collision, a context known to strongly elicit intuitive but incorrect responses. They found that even after research-based instruction [Kryjevskaia, Boudreaux, and Heins, 2014], many students who appropriately applied relevant mindware on the screening question failed to correctly answer the target question. Furthermore, students with higher CRT scores were more likely to answer the target question correctly, and to support their answer with correct and complete reasoning. Wood et al. found positive correlations between CRT scores and student performance on the Force Concept Inventory (FCI), both before and after instruction in an introductory, calculus-based mechanics course [Wood, Galloway, and Hardy, 2016]. These and other studies provide evidence of a relationship between a reasoner's tendency toward cognitive reflection and performance in physics [Kryjevskaia et al., 2019].

5.2.3. Mindware

In this study, we use responses on the Force and Motion Conceptual Evaluation (FMCE) to investigate the impact of explicit classroom discussions of the dual nature of human cognition on student learning of mechanics. The FMCE is a research-validated instrument widely used as a measure of understanding of Newtonian mechanics [Thornton and Sokoloff, 1998; Ramlo, 2008]. Results from the FMCE given before instruction in college-level introductory physics courses show that many students often apply intuitively appealing (but non-normative) ideas to describe motion. For instance, students often associate ongoing motion with an ongoing force, rather than associating force with changes in motion (i.e., acceleration). This may stem from everyday experiences in which friction forces cannot be ignored. Importantly, many students give these types of answers even after course instruction involving the application of Newton's laws to a variety of problems. In this study, we use the FMCE as a measure of student ability to recognize reasoning red flags, consider alternatives, and override intuitively appealing responses by applying formal knowledge of Newton's laws. Such reasoning pathways, involving engagement of the analytic process, are similar to those required for success on the CRT. As such, we take up two main research questions along with three tertiary questions:

1) Does explicit instruction on the dual nature of human reasoning lead to improved performance on questions that elicit intuitively appealing but incorrect ideas (i.e., questions on the FMCE)?

2) Is there a relationship between performance on the FMCE and on the CRT, and does explicit instruction on the dual nature of reasoning alter the strength of this relationship?

3) Examining the use of different CRT versions for measuring the relationship between physics performance and the tendency toward cognitive reflection?

4) Does the student CRT performance improve on the post-test? If so, a) does the level of improvement depend on the mode of instruction and b) does the change in CRT scores affect the relationships between performance on FMCE and tendency toward cognitive reflection?

5) Is the relationship between FMCE and the tendency toward cognitive reflection stable under the shifts in the CRT pre/post scores?

5.3. Methods

5.3.1. Research Context and Student Population

This study was conducted in the introductory calculus-based mechanics course at a comprehensive, Masters granting university in the northwestern region of the United States. This course is the first in a three-quarter, calculus-based introductory physics sequence, and is required for a variety of STEM majors, including chemistry, engineering, and molecular biology. About 10% of students enrolled in the course are physics majors or geology majors in the geophysics track shared by the physics and geology departments. The course includes four hours of lecture per week and a required two-hour lab but does not include a recitation section. The course is taught in lecture sections of up to 60 students each; students from different lectures mix in lab sections of 24 students. Labs are taught by undergraduate teaching assistants and use a guided inquiry curriculum focused on developing and applying conceptual and semi-quantitative models. Many of the lab activities are based on published, research-based curricula and course materials, including Tutorials in Introductory Physics [McDermott et al., 2002], Physics and Everyday Thinking [Goldberg, Otero, and Robinson, 2008], and context-rich problems developed at the University of Minnesota [Heller and Hollabaugh, 1992]. The intervention itself, described in detail below, occurred solely in the lecture section. The lecture sections meet in a standard university lecture hall, with stadium seating including fixed tablet-arm desks.

The data was collected over the course of a single academic quarter, from five lecture sections of the course taught by two instructors (A and B). Each instructor taught one section using the intervention protocol (treatment section), and the other section(s) using a modified protocol that did not make any reference to dual-process theories (control section(s)). Both instructors used active learning methodologies in their classrooms. Instructor A taught two sections, one treatment (N=48), and one control (N=44), and Instructor B taught three sections, one treatment (N=48), and two control (N=42, N=40).

Because students could register for any lab section, regardless of their lecture time, interactions between students from different lecture sections may have occurred. As discussed, the lab curriculum was adapted from research-based materials. We have found that independent of the lecture instructor, student gains on validated conceptual instruments, such as the Force Concept Inventory, are higher than nationally reported results for traditionally taught calculus-based mechanics courses [Hake, 1998; Crouch and Mazur, 2001; Cummings et al., 1999]. For example, normalized FCI gains are reliably in the range of 0.4-0.7 [Kreutzer and Boudreaux, 2012]. The overall course can thus be regarded as “interactive engagement” and “medium-high gain,” independent of the specific lecture instructor.

5.3.2. Study Design

As mentioned, each instructor taught one section using a targeted intervention focused on dual-process theories and the other section(s) using a modified protocol that did not refer to dual-process theory. The modified protocol served as a control, with roughly equal time spent on the same physics content and conceptual questions, but without framing reasoning using ideas from

DPToR. Below, we first describe the intervention protocol in detail and then summarize how the control differed.

5.3.3. Intervention Protocol Used in Treatment Groups

The intervention used in this study was intended to improve student performance on conceptual physics questions by disseminating knowledge of domain-general aspects of human reasoning and decision-making described by dual-process theories of cognition. Four separate "mini-lectures" were presented to students, in weeks 1, 3, 5, and 8 of the quarter. The course instructor gave these mini lectures using slides prepared by one of the researchers (AB). One or two days before the mini-lecture, the instructor and researcher met to review and discuss the content of the mini-lecture. In some cases, minor changes were made to the lecture slides in response to the instructor's suggestions.

The mini-lectures occurred at the beginning of a scheduled class period and lasted about 10 minutes. The first mini lecture was longer than the others and lasted about 20 minutes. The mini-lectures included 4-8 slides. Some slides presented information about dual-process theories; others had interactive prompts for students to think individually and discuss ideas with peers. The interactive portion of each mini-lecture was intended to provide a structured opportunity for students to first engage in cognitive work around a physics question involving content already covered in the course, and then to reflect on the nature of their cognition through the lens of DPToR. Students were presented with a short, conceptual physics question, and asked to answer it individually. The questions were drawn from physics education research literature. A primary selection criterion was a documented tendency of the question to elicit an appealing yet incorrect answer when administered in introductory physics courses. After responding individually to the question, students discussed their thinking with a neighbor and were guided to a normative physics explanation in a short whole-class discussion. Finally, students were asked to reflect on whether they had answered the question using “an intuitive process” (i.e., Process 1, or the “fast thinking” process) or an “analytic process” (i.e., Process 2, or the “slow thinking” process). The mini-lectures are described in more detail below.

5.3.3.1. First Mini-Lecture: Introduction to Dual-Process Theories

The goal of this mini-lecture was to present a basic framework for dual process theories of reasoning and make connections between dual process theories and the classroom learning of physics. The first few slides shared the “Ball and bat” question from the Cognitive Reflection Test [Frederick, 2005]. Students were told that this question has been given to many people, in many different settings, and that it is common for people, when first encountering this question, to confidently and quickly answer that the ball costs 10 cents. Students were guided to see that this answer is, in fact, incorrect. The mini-lecture emphasized that the primary knowledge and skills needed to complete the ball and bat task are nothing more than basic arithmetic – knowledge that all students, even those who answer the question incorrectly, can reasonably be assumed to “possess.” The fact that many reasoners initially answer the ball and bat question incorrectly was framed as an “interesting phenomenon” related to human reasoning and motivated the presentation of a simple model of cognition based on dual-process theories of reasoning. This presentation, which took up the next several slides of the mini-lecture, was drawn from Part I of the popular book “Thinking, Fast and Slow” by Daniel Kahneman [Kahneman 2011].

Students were told that human reasoning involves two distinct processes: a fast-thinking, intuitive process, and a slow, analytical process. The fast process is automatic, quick, and operates continuously with little sense of effort or voluntary control. This process was referred to in the mini-lecture as the “lizard brain.” Students were told that the fast process is involved when we respond to a question like “What is 2 + 2?” Students were then told that the slow process is deliberate, effortful, and orderly, and is involved, for example, when we review a complex argument. The slow process was described as necessarily “lazy” to accommodate the evolutionary drive to survive. That is, the slow process only comes online in certain situations that present cognitive unease. These include instances in which the fast-thinking process detects danger, encounters an impasse of some kind, red flags about the validity of the outcome of process 1 have been raised, or the reasoner has been prompted to provide an explicit justification for the validity of a decision or judgment. This presentation included the representation shown in Figure 5.2.

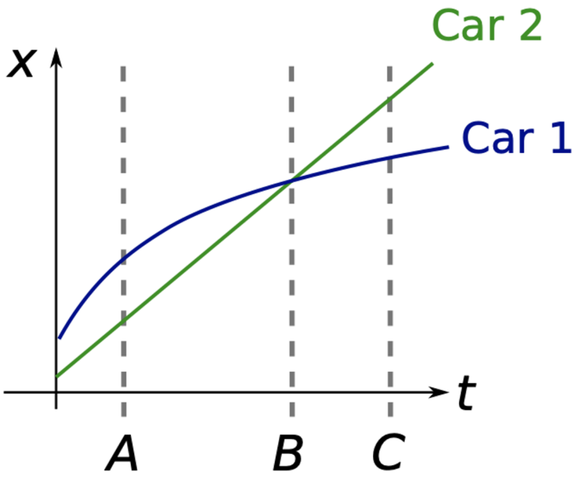


Figure 5.2. Question discussed in the first mini lecture, which typically elicits an intuitively appealing but incorrect response.

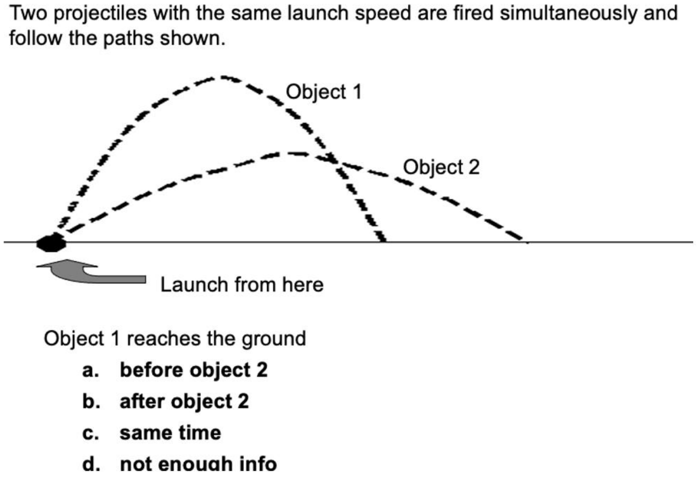
After this overview of dual-process theories, students were asked to respond individually, using “A-B-C-D cards,” to the physics question shown in Figure 5.2. This question, challenging for many introductory physics students, is known to elicit “point-slope confusion” [Heckler, 2011; Trowbridge and McDermott, 1981]. The question can initially suggest an appealing but incorrect answer that the speeds are the same at point B. More deliberate consideration of kinematics concepts covered in the course leads to a different (and correct) answer: because the two graphs have the same slopes at instant A, the speeds are the same at A.

After answering the question individually, students were prompted to compare their answers and discuss their thinking with a peer and were then asked to re-answer the question. Finally, students were asked to reflect on whether they initially answered the question with a fast, intuitive process or a deliberate, analytical process. Students shared reflections with a peer, and then volunteers were asked to share their thinking in a brief whole-class discussion. The mini-lecture concluded by encouraging students to "slow down" their thinking when encountering novel conceptual questions and check any initially appealing answers against concepts they have learned in the course.

5.3.3.2. Second Mini-Lecture

Projectile motion. Students were shown the question in Figure 5.3. and asked to answer individually using A-B-C-D cards. This question, drawn from Mazur, presents an appealing, readily available (but incorrect) response that the two projectiles reach the ground at the same time [Mazur, 1997]. A more deliberate analysis, involving a comparison of the vertical velocity components, yields a correct response that Object 1 reaches the ground after Object 2. The course had covered relevant content on projectile motion before the second mini-lecture was presented. Therefore, students had been exposed to relevant mindware that could be used to answer the projectiles question.

After answering the question individually, students were prompted to compare their answers and discuss their thinking with each other and were then asked to re-answer the question. The instructor then summarized a normative explanation. Students were then shown the diagram from Figure 5.1, presented previously in the first mini-lecture, and reminded about the "thinking fast and slow" ideas. Students were asked to reflect and speculate on whether their initial answer to the Projectiles question had been driven by fast, intuitive thinking or by more deliberate, analytical thinking. Students were further asked whether they had been aware of any influence of their "lizard brain," and/or of any moment involving engagement of their slow thinking process.


Two projectiles with the same launch speed are fired simultaneously and follow the paths shown.

Object 1 reaches the ground

a. **before object 2**
b. **after object 2**
c. **same time**
d. **not enough info**

Figure 5.3. Question discussed in the second mini lecture, which typically elicits an intuitively appealing but incorrect response.

5.3.3.3. Third and Fourth Mini-Lectures

The structure of the second mini-lecture was repeated for the third and fourth mini-lectures, but with a focus on the physics content covered at that time, the static friction force and conservation of energy, respectively. In each, students were presented with a short, conceptual question that suggested an appealing response. After answering individually, students were given opportunities to discuss their thinking and to reflect on any roles that "fast thinking" or "slow thinking" may have played in their response. In both cases, emphasis was placed on the potential value of using slow, step-by-step analysis to check any appealing answers that spring to mind upon first seeing the question.

5.3.4. Control Protocol

The protocol used in the control condition followed the basic structure and physics content of the intervention protocol, with several modifications. As in the treatment, the control involved four minilectures, administered in weeks 1, 3, 5, and 8 of the quarter. Each mini-lecture included the same class poll question as the corresponding mini-lecture for the intervention protocol but did not include any presentation of information about dual-process theories of reasoning. For the control protocol, the clicker questions were administered in a similar fashion as the intervention protocol: students were first asked to respond individually, then asked to compare their answers and discuss their thinking with each other, and then finally, a short whole-class discussion was used to go over the normative response. For the first mini-lecture of the control protocol, the kinematics graph task shown in Figure 5.2 was followed by a second, similar task, also used as a poll question. This second task provided students with additional practice using position vs time graphs to compare the speeds of two objects. This task was intended to equalize the “time on task” with the intervention protocol, given that the control protocol did not include any reference to dual process theories. The remaining mini-lectures for the control protocol did not include any additional tasks, meaning that overall, the total time on task was somewhat less for the control protocol compared to the intervention protocol.

5.3.5. Data Collection and Analysis

The FMCE and CRT were administered as pencil-and-paper, in-class assessments. Although these assignments were required, students received credit based on effort and completion, with no penalty for any incorrect answers. Students completed both assessments during a 2-hr meeting of their lab section, which occurred on the second day of instruction in the 11-week quarter. The assessments were proctored by the teaching assistant for the lab section. Students had up to 90 minutes to complete the FMCE first and then the CRT (in that order). The FMCE was administered again, with the same format and framing, on the second-to-last day of instruction of the quarter.

FMCE scores were calculated as outlined by Thornton et al. [Thornton et al., 2009] using the template developed by Smith and Wittman [Smith and Wittmann, 2008]. As suggested by the FMCE developers, seven FMCE questions were omitted from the analysis (5, 6, 15, 33, 35, 37, and 39) [Thornton and Sokoloff, 1998; Thornton et al., 2009]. In addition, traditionally, the Energy cluster is also not included in the calculation of the FMCE total score as the instrument developers argue that that concept does not necessarily characterize student understanding of motion and forces in one dimension. As such, the maximum FMCE total score in our analysis is 33 points. However, student data from the energy cluster in addition to the 33 question scores (questions 44-47). We speculated that the intervention may have a stronger impact on performance on that cluster because the DPToR-based intervention in the treatment condition was incorporated during classroom instruction on Energy. Therefore, we conducted our analyses using data from the 43-item FMCE (as recommended by the developers to assess student understanding of motion and forces in one dimension with the maximum possible score of 33 points) and the last 4 items comprising the Energy cluster (with the maximum score of 4 points).

Independent sample t-tests were performed to compare the performance of the treatment and control groups (i.e., to compare the means of two populations) and to examine the change in performance of each population as a result of instruction. Mean scores M, standard deviations

SD, t-statistics, two-sided p-values, and confidence intervals are reported below in the Results section.

The cognitive reflection test was used to measure student disposition toward cognitive reflection. The original, three-item CRT was developed in 2005 by Frederick [Frederick, 2005]. Since then, the instrument has been used extensively in psychology and has become popular in mainstream media as well. To address the instrument's possible limitations due to familiarity and the risk of floor and ceiling effects due to the small number of items, Toplak et al developed a separate 4-question instrument [Toplak, West, and Stanovich, 2014; Toplak West, and Stanovich, 2011]. Through the instrument development and validation process, it was determined that the 4-item version could be combined with the original 3-item version to create a reliable 7-item CRT without adding unnecessary redundancy. In our study, we administered the expanded 7-item CRT. To examine the impacts of our instructional intervention, we have carried out separate data analyses for the original CRT (3 items), the 4-item CRT, and the combined 7-item CRT.

Linear regression analysis was performed to examine the degree of association between student post-test ($FMCE_{post}$) and pre-test performances ($FMCE_{pre}$), CRT score (CRT), and the mode of instruction (treatment/control, T/C):

$$FMCE_{post} = \mathrm{a} + \mathrm{b} \cdot FMCE_{pre} + \mathrm{c} \cdot CRT + \mathrm{d} \cdot T/C \tag{4}$$

Different versions of this model were used to answer the research questions posed in this study. Below, we present each version separately to showcase the logic of our investigation and to facilitate discussion of the results below.

Model 0 is intended to replicate well documented findings that $FMCE_{post}$ scores are largely predicted by $FMCE_{pre}$ scores.

$$\textbf{Model 0}: FMCE_{post} = a_0 + b_0 \cdot FMCE_{pre} \quad (5)$$

Model 1 is intended to probe the impact of the instructional condition on student $FMCE_{post}$ performance. A binary indicator (a "dummy" variable) was used to assign each student a value of 1 or 0 according to whether they were part of a treatment or control group, respectively. Model 1 predicts performance on FMCE post-instruction using both $FMCE_{pre}$ score and the Treatment/Control "dummy" variables T/C. Coefficient c_1 predicts the change in the $FMCE_{post}$ score for the students in the treatment condition compared to those in the control condition with the same FMCEpre scores.

$$\textbf{Model 1}: FMCE_{post} = a_1 + b_1 \cdot FMCE_{pre} + c_1 \cdot T/C \quad (6)$$

Model 2 and Model 3 were used to examine whether or not the tendency toward cognitive reflection, as measured by the CRT, predicts student performance on $FMCE_{post}$ (Model 2) and whether this relationship, if present, is different for the treatment and control conditions (Model 3).

$$\textbf{Model 2:}\ FMCE_{post} = a_2 + b_2 \cdot FMCE_{pre} + d_2 \cdot CRT \quad (7)$$

$$\textbf{Model 3:}\ FMCE_{post} = a_3 + b_3 \cdot FMCE_{pre} + c_3 \cdot T/C + d_3 \cdot CRT \quad (8)$$

In addition, three separate versions of Model 2 were constructed with a 3-item CRT score (CRT_3), a 4-item CRT score (CRT_4), and a combined 7-item CRT score (CRT_7) to check for the predictive power of the different versions of the Cognitive Reflection Test.

Akaike Information Criterion (AIC) model selection was applied to distinguish among models with multiple predictors. Changes in AIC are reported where appropriate.

To aid in interpreting the results, we report standardized and unstandardized coefficients. Unstandardized coefficients estimate how many units the outcome would change for every unit of change in the predictor, holding all other predictors constant. Standardized coefficients estimate the number of standard deviations the outcome changes for every standard deviation change in the predictor, holding all other predictors constant. Standardized coefficients are convenient for comparing the relative strengths of each predictor. The effect size associated with each predictor was determined by calculating f^2:

$$f^2 = r^2_{part} / 1 - r^2_{part}, \tag{9}$$

where r^2_{part} is a squared semi-partial (or "part") correlation for a predictor [47]. For a model with a single predictor, R^2 of the model is used instead of. Values $f^2 = 0.02$, $f^2 = 0.15$, and $f^2 = 0.35$ indicate small, medium, and large effect sizes, respectively.

Evaluation of the statistical assumptions indicated minor concerns about meeting established criteria for homoscedasticity and normality of residuals. Although linear regression is robust under minor violations of these assumptions (especially for large sample sizes), we reevaluated our models by performing bootstrapping to alleviate any concerns with the accuracy of confidence intervals and tests of significance. Bootstrap confidence intervals and significance values do not rely on assumptions of normality or homoscedasticity. This paper reports 95% bias-corrected and accelerated confidence intervals and standard errors based on 1000 bootstrap samples. Finally, in all models presented here, the variance inflation factor (VIF), which indicates whether a predictor has a strong linear relationship with the other predictors, does not deviate significantly from 1, suggesting that all the models meet the assumption for no multicollinearity. Similarly, bootstrapping was performed to mediate any minor violations of assumptions for t-tests.

5.4. Results

5.4.1. Establishing Equivalence of Treatment and Control Groups

First, we compare the students' background knowledge across the treatment and control groups, as measured by FMCE pre-test scores. Independent samples t-tests compared scores for both the 43-item FMCE instrument as well as the Energy cluster. No statistically significant differences in performance were observed on either measure (see Table 5.1). On the 43-item FMCE, the pre-course performance of the treatment group (M=32%, SD = 26%) was not distinguishable from that of the control group (M=30%, SD = 26%; t (214) =0.61, p=0.54). Similarly, performance on the Energy cluster questions of the treatment group (M=30%, SD = 37%) and the control group (M=29%, SD =36%) was comparable (t (214) =0.31, p=0.76). These results suggest that any differences in post-test FMCE scores are unlikely to be the result of differences in students' background knowledge of mechanics concepts.

Table 5.1. Correct response rates on the 43-item FMCE before *($FMCE_{pre}$)* and after *($FMCE_{post}$)* instruction and on the Energy cluster before *(E_{pre})* and after *(E_{post})* instruction in the treatment and control groups. The average performance of all students, independent of instructional condition, is included in the aggregated results.

Instructional condition	$FMCE_{pre}$	$FMCE_{post}$	E_{pre}	E_{post}
Treatment	32%	60%	30%	72%
Control	30%	61%	28%	73%
Aggregated	31%	61%	29%	72%

Next, we compare the two groups' tendencies toward cognitive reflection, as measured by CRT scores. The mean scores on the 3-item CRT for the treatment (M = 1.7, SD = 1.0) and control (M= 1.7, SD = 1.0) groups showed no statistically significant difference, t (214) = 0.67, p = 0.51. A similar result was obtained for the 7-item CRT (treatment group: M = 3.3, SD = 1.8; control group: M = 3.4, SD = 1.6; t (214) =0.76, p = 0.45). These results allow us to treat the two groups of students as equivalent before instruction in terms of both physics background knowledge and tendencies toward cognitive reflection.

5.4.2. Research Question 1: To What Extent, if at All, Does Explicit Instruction on Dual-Process Theories of Cognition Improve Student Performance on Physics Tasks That Tend to Elicit Intuitively Appealing but Incorrect Responses?

A comparison of FMCE scores post-instruction does not reveal any statistically significant difference in the performance of the treatment and control groups (treatment group: M=60%, SD = 28%; control group: M=61%, SD = 30%; t (214) =0.30, p =0.76). Similarly, no statistically significant difference in performance was observed on the Energy cluster (treatment group: M=72%, SD = 37%; control group: M=73%, SD = 36%; t (214) =0.33, p=0.74).

Furthermore, as expected, linear regression Model 0 yields a strong and statistically significant relationship between $FMCE_{pre}$ and $FMCE_{post}$ scores with a large effect size (f2=0.74), as shown in Table 5.2. The model estimates that as the $FMCE_{pre}$ score increases by one standard deviation, the $FMCE_{post}$ score increases by 0.65 standard deviations. Moreover, this relation does not appear to change in Model 1, which includes a dummy variable to indicate the treatment or control condition. These results suggest that explicit instruction on dual-process theories of cognition did not significantly impact students’ post-course mechanics performance.

Table 5.2. Linear models of predictors of performance on FMCE after instruction, FMCE$_{post}$. Confidence intervals (in parentheses) and standard errors are based on 1000 bootstrap samples. R^2=0.43 for model 0, R^2=0.43 for model 1, R^2=0.46 for model 2, and R^2=0.46 for model 3.

		Coefficients	Standard error	Standardized coefficients	Sig. (2-tailed) p-values
Model 0	Constant	37.56 (33.09, 42.15)	2.34	-	<0.001
	FMCE$_{pre}$	0.76 (0.67, 0.86)	0.04	0.65	<0.001
Model 1	Constant	38.73 (33.98, 43.51)	2.59	-	<0.001
	FMCE$_{pre}$	0.76 (0.68, 0.88)	0.04	0.66	<0.001
	T/C	-2.89 (-9.5, 4.01)	3.14	-0.05	0.357
Model 2	Constant	28.57 (21.7, 36.3)	3.6	-	<0.001
	FMCE$_{pre}$	0.65 (0.55, 0.76)	0.05	0.56	<0.001
	CRT7	3.65 (1.59, 5.61)	1.06	0.20	<0.001
Model 3	Constant	29.53 (21.73, 37.80)	3.91	-	<0.001
	FMCE$_{pre}$	0.65 (0.55, 0.76)	0.05	0.57	<0.001
	CRT7	3.60 (1.58, 5.54)	1.06	0.20	<0.001
	T/C	-2.03 (-8.13, 3.45))	3.10	-0.03	0.505

5.4.3. Research Question 2: Is There a Relationship Between Performance on FMCE and Tendency Towards Cognitive Reflection? If So, Does the Strength of the Relationship Depend on the Mode of Instruction?

Regression coefficients for Models 2 and 3 are presented in Table 5.2, above. Model 2 adds an additional variable, CRT7, to model 0, which improves the model fit as indicated by the increase in R^2 and the decrease in AIC between model 0 and model 2 (ΔR^2=0.03, p=<0.001; ΔAIC=-11). These results indicate that student tendency toward cognitive reflection impacts student performance on FMCE post-instruction scores even after controlling for student background knowledge (as measured by the $FMCE_{pre}$ score). Students entering a physics course with similar physics background perform better after instruction if they have a stronger tendency towards cognitive reflection, e.g., if they are more likely to mediate their intuition by engaging in analytical thinking. The relationship between CRT score and FMCE_{post} score has a small effect size (f^2=0.03), while the relationship between FMCE_{pre} score and FMCE_{post} score remains large (f^2=0.34), even after accounting for cognitive reflection.

The weaker link between $FMCE_{post}$ and CRT scores is not surprising. $FMCE_{pre}$ scores can be viewed as a proximal factor with a direct link to $FMCE_{post}$ scores. in contrast, the tendency toward cognitive reflection, is a more distal factor as its utility depends on the strength of the relative knowledge necessary to answer a given task correctly. Reasoners who possess strong cognitive reflection skills but lack relevant knowledge are not likely to arrive at a correct answer. On the other hand, cognitive reflection skills could be instrumental for engaging in error detection in cases where incorrect responses stem from the relevant knowledge being present but not accessible to a reasoner due to highly salient but non-diagnostic features of a task.

Model 3 which added the *T/C* variable to model 2, was no more accurate than model 2 ($\Delta \mathrm{R}^2$=0.001, p=0.51), indicating that the mode of instruction does not affect the relationship between FMCE and CRT scores. This result suggests that including explicit classroom discussion of dual-process theories does not boost student cognitive reflection to a level detectable by performance on the FMCE post-instruction.

A separate set of models, similar to those in Table 5.2, was used to probe the relationships between FMCE energy cluster pre-test scores, post-test scores, CRT score, and the mode of instruction. The results are shown in Table 5.3.

Table 5.3. Linear models of predictors of student performance on the FMCE energy cluster after instruction, E_{post}. Confidence intervals (reported in the parentheses) and standard errors are based on 1000 bootstrap samples. R^2=0.06 for model 0, R^2=0.06 for model 1, R^2=0.14 for model 2, and R^2=0.14 for model 3.

		Coefficients	Standard error	Standardized coefficients	Sig. (2-tailed) p-values
Model 0	Constant	65.47 (59.39, 71.16)	3.33	-	<.001
	E_{pre}	0.24 (0.14, 0.35)	0.05	0.24	<0.001
Model 1	Constant	66.33 (33.3, 44.2)	3.94		<0.001
	E_{pre}	0.24 (0.14, 0.35)	0.05	0.24	<0.001
	T/C	-2.04 (-11.67, 8.23)	5.02	-0.03	0.676
Model2	Constant	45.82 (34.98, 55.55)	5.60	-	<0.001
	E_{pre}	0.20 (0.11, 0.30)	0.05	0.20	0.002
	CRT7	6.19 (3.23, 9.14)	1.44	0.28	<.001
Model 3	Constant	46.25 (34.59, 57.15)	5.94	-	<.001
	E_{pre}	0.21 (0.11, 0.30)	0.05	0.20	0.002
	CRT7	6.17 (3.25, 9.11)	1.45	0.28	<.001
	T/C	-0.91 (-9.92, 7.79)	4.64	-0.01	0.846

Results for the energy cluster are analogous to those from the entire 43-item FMCE. In particular, the E_{pre} and *CRT7* scores are statistically significant predictors of student performance on the energy cluster post-instruction, while the mode of instruction does not appear to impact performance. Student performance on energy cluster questions improved dramatically, from ~29% on the pre-test to ~73% on the post-test. (See Table 5.2). However, students with a stronger tendency toward cognitive reflection and stronger relevant physics background knowledge seemed to benefit more from course instruction. Students who scored one standard deviation higher on *CRT7* scored 0.28 standard deviations higher on E_{post} (keeping the E_{pre} score constant). Similarly, students who scored one standard deviation higher on E_{pre} scored 0.20 standard deviations higher on E_{post} (keeping the *CRT7* score constant). The effect size of the *CRT7* score is in the small-to-medium range, f_{CRT}^2=0.08; while the effect size of the E_{pre} score is small (f_{Epre}^2=0.04). We note that in Model 2 for the 43-item FMCE, the effect size of the $FMCE_{pre}$ score is much higher, f_{FMCEpre}^2=0.34. This suggests that student performance on the energy cluster after instruction is less dependent on performance before instruction, whereas pre- and post- performance for the 43-item instrument are greatly related.

Overall, we find that a classroom discussion of the role of dual-process theory in learning and reasoning about energy concepts (i.e., as presented to the treatment group in the fourth mini-lecture) did not boost student performance to a level detectable in our study.

5.4.4. Research Question 3: Examining the Use of Different CRT Versions for Measuring the Relationship Between Physics Performance and the Tendency Toward Cognitive Reflection

As discussed, our study made use of the extended version of the Cognitive Reflection Test. This allows us to examine the extent to which different versions of the CRT provide consistent measures of the tendency toward cognitive reflection in the context of the learning and teaching of physics. In this section, we compare three different versions of Model 2, using *CRT3, CRT4,* or *CRT7,* as a predictor of performance on $FMCE_{post}$. Variables *CRT3, CRT4, and CRT7* represent scores on the original 3-item version of the CRT, the four items added to the original version of the CRT, and the entire extended 7-item version of the CRT, respectively. All models have improved predictive power over Model 0 which only includes $FMCE_{pre}$ scores.

As is evident from Table 5.4, all models show nearly identical relationships between student tendencies toward cognitive reflection and the performance on FMCE post-instruction, even after accounting for *$FMCE_{pre}$* score. This result suggests that different versions of the CRT provide comparable measures of the tendency toward cognitive reflection, at least in relation to performance on the FMCE.

Table 5.4. Linear regression models of predictors of student performance on the FMCE post-instruction, *$FMCE_{post}$*. Confidence intervals (reported in the parentheses) and standard errors are based on 1000 bootstrap samples. R^2=0.45, p=0.003, $f^2_{FMCEpre}$=0.39, f^2_{CRT3}=0.02 for model 2 with CRT3; R^2=0.45, p=0.005, $f^2_{FMCEpre}$=0.48, f^2_{CRT4}=0.02 for model 2 with CRT4; R^2=0.46, p<0.001, $f_{FMCEpre}{}^2$=0.34, $f_{CRT7}{}^2$=0.03 for model 2 with CRT7.

		Coefficients	Standard error	Standardized coefficients	Sig. (2-tailed) p-values
Model 2 with CRT3	Constant	31.59	3.07	-	<.001
	$FMCE_{pre}$	0.68	0.05	0.59	<.001
	CRT3	4.92	1.66	0.17	0.003
Model 2 with CRT4	Constant	31.72 (25.03, 38.76)	3.16	-	<.001
	$FMCE_{pre}$	0.70 (0.59, 0.80)	0.06	0.60	<.001
	CRT4	4.68 (1.40, 8.12)	1.67	0.15	<.001
Model 2 with CRT7	Constant	28.57 (21.7, 36.3)	3.6	-	<0.001
	$FMCE_{pre}$	0.65 (0.55, 0.76)	0.05	0.56	<0.001
	CRT7	3.65 (1.59, 5.61)	1.06	0.20	<0.001

5.4.5. Research Question 4: Does Student CRT Performance Improve on the Post-Test? If So, a) Does the Level of Improvement Depend on the Mode of Instruction and b) Does the Change in CRT Scores Affect the Relationships Between Performance on FMCE and Tendency Toward Cognitive Reflection?

In this study, the CRT was administered to all students twice, at the beginning and end of the course. First, we probed whether any differences in pre/post-CRT scores could be detected for all students in the study (CRT data was aggregated independent of the mode of instruction). Then, we explored whether this difference could be attributed to the mode of instruction.

Paired t-tests were performed to compare pre- and post-course scores on CRT3, CRT4, and CRT7. The average CRT3 score changed from 1.71 to 2.12, an increase of ~0.41 points (~14%). This difference is statistically significant (p<0.001), with a medium effect size (Cohen's d=0.49). The average CRT4 score changed from 2.27 to 2.45, an increase of about 5%. While this difference is statistically significant (p<0.02), the effect size is small (Cohen's d=0.17). The average CRT score for the entire 7-item instrument changed from 3.98 to 4.57, an increase of about 8%. This difference is statistically significant (p=0.001) with a medium effect size (Cohen's d=0.42). The results suggest that a CRT score increases on average with small-to-medium effect size, regardless of which version of the CRT is used.

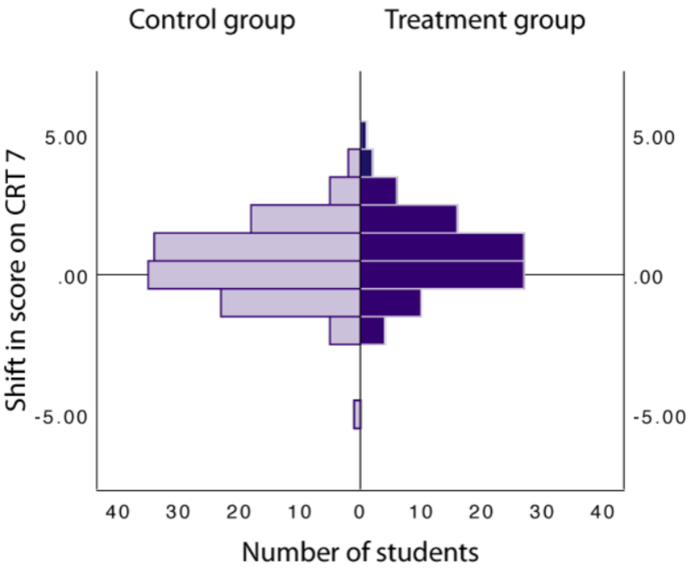


Figure 5.4. Distributions of shift scores on CRT7 for treatment and control groups.

Pre- and post-course CRT response data was used to generate a "shift score" for each student on each CRT item. We used "1" to represent a positive shift (from incorrect on the pretest to correct on the posttest), 0 to represent no shift, and -1 to represent a negative shift. The Mann-Whitney U test (a nonparametric test that compares the rank order of scores between two independent samples) was applied to explore the difference in the distributions of shift scores for the students in the treatment and control groups. We found that the distributions of shift scores differ for the treatment and control groups for the *Bat-and-ball* CRT item (Mann-Whitney U=6463.5, p=0.037, with a small effect size r=0.14), but not for any other CRT item. (Recall that the *Bat-and-ball* item was used explicitly in instruction in the treatment condition during the first mini-lecture.) The average performance on the *Bat-and-Ball* item increased by ~12% (from 56.9% to 69.1%) for the control group and by ~26% (from 60.2% to 86%) for the treatment group. The improved performance on this question for the control group is consistent with the average small improvement in performance on all other CRT items not explicitly discussed during instruction.

Similar analyses for the three sets of CRT questions (i.e., CRT3, CRT4, and CRT7) reveal no statistically significant difference in the distribution of shift scores between the treatment and control groups (Mann-Whitney U=6262.0, p=0.19; U=6121.0, p=0.35; U=6421.5, p=0.11, respectively). Figure 5.4. illustrates distributions of shift scores for the two groups of students.

The results suggest that the average CRT score appears to increase after instruction. While the average scores of all versions of the CRT increase, student individual performance shifts in both directions with a slightly larger number of students moving toward correct answers, as shown in Figure 5.4. for CRT7. However, there is no statistically significant difference between the distributions of shifts in student scores in the treatment and control groups.

5.4.6. Research Question 5: Is the Relationship Between FMCE and the Tendency Toward Cognitive Reflection Stable Under the Shifts in CRT Pre/Post Scores?

A statistical analysis similar to that discussed in Research Question 3 was performed. Table 5.5 shows the results from model 2 with variables $CRT3_{post}$, $CRT4_{post}$, and $CRT7_{post}$ corresponding to student scores obtained after post-instruction administration of the CRT.

Table 5.5. Linear regression model 2 for the performance on the FMCE post-instruction, *$FMCE_{post}$* (R^2=0.46, p<0.001, $f^2_{FMCEpre}$=0.42, $f^2_{CR42,ost}$=0.03 for Model 2 with CRT_{3_post}; R^2=0.47, p<0.001, $f^2_{FMCEpre}$=0.42 , $f^2_{CRT4\_post}$=0.04 for Model 2 with CRT_{4_post}; R^2=0.47, p<0.001, $f^2_{FMCEpre}$=0.46, $f^2_{CRT3\_post}$=0.06 for Model 2 with CRT_{7_post})

		Coefficients	*Standard error*	*Standardized coefficients*	*Sig. (2-tailed) p-values*
Model 2 with $CRT3_{post}$	*Constant*	31.59	3.07	-	<.001
	$FMCE_{pre}$	0.68	0.05	0.59	<.001
	$CRT3_{post}$	4.92	1.66	0.17	0.003
Model 2 with $CRT4_{post}$	Constant	31.72 (25.03, 38.76)	3.16	-	<.001
	$FMCE_{pre}$	0.70 (0.59, 0.80)	0.06	0.60	<.001
	$CRT4_{post}$	4.68 (1.40, 8.12)	1.67	0.15	<.001
Model 2 with $CRT7_{post}$	*Constant*	28.57 (21.7, 36.3)	3.6	-	<0.001
	$FMCE_{pre}$	0.65 (0.55, 0.76)	0.05	0.56	<0.001
	$CRT7_{post}$	3.65 (1.59, 5.61)	1.06	0.20	<0.001

As is evident from Tables 5.4 and 5.5, the relationships among FMCEpost, $FMCE_{pre}$, and CRT are nearly identical independent of the version of CRT or the timing of its administration: $FMCE_{pre}$ is linked with the $FMCE_{post}$ with a large effect size while tendency toward cognitive reflection has small-to-medium effect side on $FMCE_{post}$. This result suggests that the observed relationships are stable under test/retest conditions.

5.5. Discussion and Conclusion

Prior research suggests that even students with the knowledge and skills necessary to answer physics tasks correctly often default to accepting an intuitively appealing but incorrect response without further scrutiny. As discussed, dual process models of reasoning provide a mechanism that can account for this phenomenon. Our teaching experience shows that the interplay of fast and slow thinking is of interest to many students. Furthermore, discussion of this mechanism may lessen the stigma of incorrect answering by attributing some incorrect answers to universal aspects of humans processing and decision making.

In this paper, we have argued that a reasonable first step in mitigating the phenomenon described above is to promote student awareness of dual process theories of reasoning, including how such theories relate to classroom performance in physics. If students become aware of reasoning hazards inherent in human reasoning (as described by DPToR), such as the path of cognitive frugality, they may be motivated to learn and practice strategies to recognize and override these ubiquitous hazards. In this study, we administered a targeted instructional intervention involving explicit classroom discussions of dual-process theories of reasoning in a introductory mechanics course. The goal was to enhance student knowledge of domain-general aspects of human cognition and emphasize the role of cognitive reflection in mediating appealing, first-available responses that may not be normative. We examined whether this type of explicit instruction, which supplemented other student-centered, research-based teaching strategies, could improve student performance on tasks that are known to elicit intuitively appealing but incorrect responses, such as those included in the FMCE.

Two instructors participated in the study. Each instructor taught one section in which they implemented the intervention (i.e., a treatment condition), as well as at least one other section in which they used a modified protocol that did not reference dual-process theories (i.e., a control condition). Results suggest that conducting explicit classroom discussions of DPToR was insufficient for improving student performance on challenging qualitative physics questions that elicit strong, intuitively appealing incorrect responses. We also showed a link between student performance on the FMCE and student tendency toward cognitive reflection. Namely, after controlling FMCE_{pre} score, students with a higher tendency toward mediating intuitive responses by engaging in analytical thinking were more likely to answer FMCE questions correctly on the post-test. We did not, however, detect any difference in the relationship between CRT and $\text{FMCE}_{\text{post}}$ scores between the treatment and control groups. Our data revealed that the relationship between performance on the FMCE and the CRT appears to be independent of the version of the cognitive reflection test. Linear regression models predict similar relationships for the 3-item CRT, the additional four items on the expanded version of the CRT, and the entire 7-item expanded CRT, which includes both the original three items and the additional four items. Interestingly our results showed a statistically significant increase in average CRT scores, with small-to-medium effect sizes, regardless of the version of the CRT used. Examining whether students' shifts were positive or negative showed that the distributions of "shift scores" did not differ between the treatment and control groups, except for the Bat-and-ball CRT item, which was explicitly discussed in the treatment condition. However, there was no significant difference in the distribution of shift scores for other CRT items between the two groups. While average CRT scores increased after instruction, individual student performance varied, with a slightly larger number of students moving toward correct answers. The relationship between CRT score and $FMCE_{post}$score did not change, despite the change in CRT score, so while CRT scores may change over time it seems that that the relationship between conceptual understanding, as measured by the FMCE, and cognitive reflection, as measured by the CRT, is stable over time.

We propose two possible interpretations of the null results of this study. First, the study was conducted in a learning environment that already utilized student-centered techniques, resulting in higher learning gains than those reported nationally for traditional instruction [Hake, 1999; Von Korff et al., 2016]. As such, the potential impacts of explicit instruction on the dual nature of reasoning may have been overshadowed by significant gains due to other student-centered instruction. Although it may be easier to detect the effect of the intervention on FMCE performance in a more traditional environment, we do not necessarily advocate for such a study since student-centered STEM instruction should become a norm. At the same time, we recognize that traditional lecture remains a prevalent mode of instruction throughout the undergraduate STEM curriculum. Therefore, probing the impact of the fairly easy-to-implement DPToR intervention described here may be worth pursuing. Furthermore, our extensive experience in sharing research findings on student cognition from the perspective of DPToR indicates that STEM faculty, including physics educators, are generally curious and receptive to learning about the dualistic nature of human reasoning and how it may impact the learning and teaching of physics. Discussing research informed by DPToR with instructors may help promote incremental instructional changes and generate data on STEM teaching in diverse learning environments.

Second, although no effect of the intervention on student FMCE performance was detected in this study, we recognize that other important aspects of student learning may have been affected. For instance, student attitudes toward learning and epistemology are generally difficult to shift in a positive direction. The tendency of evidence-based practices to encourage students to confront their own incorrect answers often leads to negative attitudes towards those practices, at least by those unfamiliar with them. It is possible that explicit discussions of universal factors of human reasoning could mediate some of those negative feelings and improve student attitudes toward learning in physics (and beyond). It is critical to help students realize that mistakes are unavoidable, normal, and present opportunities to grow. Making and correcting mistakes is a natural part of learning, not a sign of inadequacy or failure. Finally, developing effective reasoning strategies in a physics context can have far-reaching benefit, as the metacognitive skills of recognizing reasoning red flags and knowing how to validate or reject a response are associated with expert thinking in any field.

6. CONCLUSION

This dissertation aims to use theoretical frameworks from cognitive psychology to investigate aspects of teaching and learning that are particularly resistant to change. For teaching and learning to be successful, practitioners must be able to use the most effective instructional strategies. As such, education researchers must identify t challenging aspects of teaching and learning and address them practically and effectively through research and practice. T The development of evidence-based instructional strategies (EBIPs) can have a significant impact, but only if STEM instructors are willing to implement these strategies, and, even then, only if they are supported in implementing them effectively. This dissertation focused on both aspects: 1) identifying and addressing student reasoning difficulties in physics and 2) assessing the effectiveness of the professional development of STEM instructors. Specifically, The Theory of Planned Behavior (TPB) is used to understand whether a behavior, such as implementing EBIPs, will be performed. The Dual Process Theories of Reasoning and Decision-Making (DPToR) are used to understand student reasoning patterns. Investigations include the development and validation of a research instrument to assess professional development programs and three targeted classroom interventions to improve student reasoning.

Education research over the past few decades has shown that evidence-based instructional practices (EBIPs) are more effective than traditional lecture methods, which often entail the passive transmission of information. EBIPs can significantly impact student learning, including improved performance on concept inventories, decreased achievement gaps, greater self-efficacy and retention, improved attitudes towards learning, increased student engagement, improved sense of belonging, and more productive epistemological beliefs.

Despite the proven benefits of active learning methodologies (ALS), many factors have slowed their widespread implementation. While the use of EBIPs has increased, they are not yet the norm. Research has identified obstacles to implementing instructional innovations, and efforts have been made to understand and address these obstacles, particularly through professional development opportunities for college instructors. However, systematic assessment of these efforts remains challenging.

Chapter 2 focuses on a fairly underdeveloped aspect of STEM Education research, professional development assessment. While research has shown [Henderson et al 2011; Khatri et al 2015; Borrego and Henderson 2014; Henderson 2008] that large scale educational change necessitates concerted, supported, and intentionally designed change efforts such as professional development, the tools to assess professional development in STEM Ed has room for improvement [Chasteen et al 2016; Chasteen and Chattergoon 2020]. This is not the case in other fields where the Theory of Planned Behavior, a theoretical framework that explains intentions towards behavior in terms of 3 encompassing factors of perception, has been used in other fields to assess behavioral changes [Armitage and Conner, 2001]. Additionally, this chapter discusses the phenomenon of response shift bias, in which participants' inexperience prior to an intervention (or significant growth during an intervention) causes them to effectively understand pre- and post- test self-report surveys differently, potentially invalidating a fundamental tenant of traditional pre-post methodologies.

The Theory of Planned Behavior was used to develop an instrument which assesses respondents' beliefs and intentions about active learning strategies, the Beliefs and Intentions to Use Active Learning Strategies Survey (BIUALS). Aroon et al.'s framework for instrument validation is used to validate the BIUALS, discussing internal consistency, temporal stability, and relation to other known variables. The analysis presented provides evidence for the validity of the BIUALS as well as produces two other points of interest and justifies the following section, which discusses an analysis of some BIUALS data. Interestingly, factor analysis of survey items supports two interpretations of the norms category, the use of norms as intended, and a potential split of norms into two categories: a department's stated support of active learning strategies and the participant's perception of the department's support. This indicates a potential mismatch between the words and the actions of a department towards instructional change. These validation efforts also demonstrate that retrospective-pretest scores are stable over time, providing evidence for the validity of the retrospective pre-test methodology.

Given the validity of the BIUALS, some initial data are analyzed and discussed. The results of this analysis show that traditional pre-post assessments would have yielded null results, suggesting minimal effects of the PD program. However, the retrospective pre-test reveals much more optimistic results, indicating significant changes in participants' beliefs and intentions towards ALS. The difference between pre-PD scores and retrospective scores supports the claim that response shift bias was an issue. The existence of response shift bias and the stability of retrospective scores over time provide further evidence for the use of the retrospective pretest design for future self-report assessments in STEM Ed or PER. Beyond these far-reaching claims this analysis also suggests that Gateways ND positively impacted participants, especially in terms of attitudes and perceived behavioral control, which traditional pre-post methodology would have missed.

Beyond survey and methodological validity, this study also explores the relationships between intentions and the factors described by the TPB. Perceived behavioral control, e.g. beliefs about self-efficacy and perception of ease of use and barriers to use, emerges as the strongest predictor of intentions, highlighting the importance of increasing faculty confidence and control over their classrooms to promote the use of ALS. Despite possibly delineating into two categories, norms seemed to have no impact on intentions to implement ALS. Future work should investigate the generalizability of these relationships and examine the difference between various populations of instructors.

Overall, the study underscores the utility of the TPB and the retrospective pre-test methodology in evaluating PD programs and highlights the potential for interdisciplinary approaches to enhance assessments in STEM education research, as well as providing an optimistic assessment of Gateways-ND's effect on its participants and providing interesting insights into NDSU faculty motivations.

Chapter 3 begins a three-chapter examination of inconsistent student reasoning, wherein students demonstrate correct reasoning in one context but fail to apply that same reasoning in conceptually identical contexts. The following chapters use the Dual Process Theories of Reasoning as a lens to understand how these reasoning inconsistencies occur, and why they are often so persistent, arguing that a fundamental aspect of human reasoning explains this tendency. DPToR describes reasoning as the interaction between two processes, process 1, a fast, subconscious, heuristic process that produces mental models, and process 2, a slow, deliberate, analytic process that assesses the mental models developed by process 1. The inconsistent reasoning phenomenon is explained as the result of 1) distracting context clues, which cue the development of incorrect initial mental models, and 2) a failure by process 2 to reject the incorrect model. Process 2 can fail in several ways, most simply by not activating at all (cognitive miserliness) but may also activate and fail to reject the model due to overconfidence in the incorrect mental model [Johnson-Laird, 2006; Toplak, West, and Stanovich, 2011;

Thompson, 2011], due to its own reasoning biases (such as confirmation bias) [Nickerson 1998], or due to the weakness (or absence) of the relevant knowledge and skills.

Chapter 3 explores one way of improving student reasoning performance in physics. Namely, the use of collaborative exams to create a high-stakes environment in which process 2 is modeled through the process of socially mediated metacognition [Goos et al., 2002; Vygotsky and Cole, 1978; Shirouzu et al., 2002; Siegel, 2011]. In a high-stakes exam environment where both answer and reasoning are graded, students should be more motivated to present correct reasoning, leading students to examine their thinking rather than simply accepting an answer as correct. If successful, students will either engage process 2 and overcome intuitively appealing but incorrect first mental models or be exposed to other students process 2 engagement. The collaborative exam treatment was compared against a control condition, in which a review session was substituted for the collaborative exam, providing a more traditional instructional style with equivalent time on task to compare against.

Several sets of Screening-Target questions were included across midterm exams, and target questions were again included on the final exam to determine the long-term effects each instructional condition had on student reasoning. Analysis of student responses reveals that both the instructor lead review (control) population and the collaborative exam (treatment) population experienced similar significant improvements in performance on the final exams. Despite comparable improvements, we still advocate for the use and testing of the collaborative group exam approach, especially in courses where group work is common. Collaborative exams offer additional benefits beyond performance improvements without, according to our data, negatively impacting student learning gains. These benefits may include reduced test anxiety, fostering social networks of support among students, and aligning assessment with instructional strategies (group work) [Wieman, Rieger, and Heiner, 2014; Rieger and Heiner, 2014; Zimbardo, Butler, and Wolfe, 2003]. Despite the null results of the presented study, student performance on the final in both conditions shows that collaborative exams can still serve as successful instructional tools, and a growing body of evidence shows they can have effects outside of conceptual gains. The potential benefits for student learning and students' attitudes towards learning should encourage future evaluation of the benefits and best practices of collaborative exams.

Chapter 4 continues the examination of inconsistent reasoning by evaluating the impact of a three-stage instructional intervention focused on the application of Newton's 2nd law to objects at rest. Two conceptually analogous questions were posed to students, a screening question regarding a block at rest with a force applied, and a target question involving a magnet on a fridge with an upwards force from a hand. Pre-intervention analysis shows that, as expected, students were relatively successful when responding to the screening question (~50% correct) but were much less successful when responding to the target question (~20% correct), and students that failed to answer the screening question were extremely unlikely to answer the target question correctly (~4% of students who incorrectly answered the block question).

The three stages of the intervention were designed through the lens of DPToR and aimed to improve both students' mindware (conceptual understanding) and to engage their cognitive reflection, hopefully facilitating more consistent and successful process 2 intervention. This was done by 1) asking students individually to reconsider the block and magnet question, with the intention of encouraging students to connect the two conceptually; 2) having students work in pairs to reconsider both questions, prompting socially mediated metacognition to occur, ideally allowing students to train/improve their reasoning in this context; and 3) asking students to work through guiding questions focused on considering similar scenarios to the magnet question, highlighting the concepts involved and drawing students away from intuitively appealing responses. With the block question serving as the primary measure of mindware, the Cognitive Reflection Test [Frederick, 2005] was used to measure students' cognitive reflection.

Analysis of student responses showed that each intervention stage led to a 10%-30% improvement in student performance on the target question, despite salient distracting cues. Most students demonstrate correct reasoning at some point during instruction (74%-90% success rate). However, a significant portion of students (30%) reverted to incorrect reasoning on a comparable (but not identical) question included on the course exam despite responding to the target question correctly during instruction. These results underline the persistent nature of student inconsistent reasoning patterns and highlight the power intuitive thinking has to overshadow formal knowledge learned in class.

Logistic regression models indicated that both mindware and cognitive reflection skills predicted student performance after the intervention, but not before it. This indicates that instruction focused on improving cognitive reflection, in the absence of sufficient mindware, is unlikely to help most students. Specifically, the subset of students that did not experience stage 1 or 2 of the intervention showed no link between CRT score and posttest performance in the regression models. This supports both the claim that sufficient mindware is necessary for cognitive reflection to play any role, and that the later stages of the intervention meaningfully improved mindware and engaged cognitive reflection skills. For students that experienced all 3 stages of the interventions, those with both sufficient mindware and high cognitive reflection skills were significantly more likely to answer the post-test question correctly, emphasizing the importance of taking both aspects of reasoning into consideration.

Chapter 5 finishes the examination of inconsistent student reasoning, focusing on the conscious and subconscious differences between process 2 and process 1. DPToR can explain inconsistent student reasoning as a failure of process 2 to engage when process 1 creates an incorrect mental model. This process happens subconsciously, as students may create and accept a mental model without ever checking that model for consistency or validity. We argue that students are unlikely to be aware of this process as they reason through physics questions and hypothesize that teaching students how they think might encourage them to intentionally engage process 2 more frequently. An intervention consisting of 4 mini-lectures was developed to teach students about their reasoning through the lens of DPToR and model how students might accept an intuitively appealing first mental model without evaluating it, as well as how to productively engage with those models. This intervention is compared against a control condition with comparable time on task but without reference to DPToR,

The Force and Motion Conceptual Evaluation (FMCE) was used to assess the impact of the interventions on student performance, due to its focus on questions which tend to elicit intuitively appealing incorrect responses. The Cognitive Reflection Test was also used to gauge students' tendency to cognitively reflect, and the opportunity was taken to test several versions of the CRT, as well as their stability over time.

Analysis of the data found no statistically significant differences between the treatment condition and the control condition, which used similar mini lectures but made no reference to DPToR. While no differences between the control and treatment were found, conceptual gains in both groups were relatively large ($< g > \sim 0.4$). This study was conducted in classes which make frequent use of active learning strategies, and specifically strategies which encourage students to consider and reflect on their initial responses. It is possible that the effects of specific

DPToR instruction were overshadowed by the effects of learning strategies with a similar focus of improving student reasoning and reflection.

While the intervention itself did not produce distinguishable results, this study did provide useful data regarding the Cognitive Reflection Test. A statistically significant link was established between CRT score and the FMCE, and this link was shown to be stable over CRT retest. Multiple versions of the CRT were administered, and examination of their results show that each version functions effectively identically in relation to FMCE scores. This study also found that for both treatment and control populations CRT scores were not stable over time. Student responses shifted in both directions, indicating that retest was not an overwhelmingly consistent issue, but it is unclear if these shifts are due primarily to variance or changes in student cognitive reflections abilities.

This work is based on work supported by the National Science Foundation under Grant Nos. DUE-1525056, DUE-1821390, DUE-1821123, DUE-1821400, DUE-1821511, DUE-1821561, DUE-1431940, DUE-1431541, DUE-1431857, DUE-1432052, and DUE-1432765.

REFERENCES

Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical review special topics-physics education research, 2*(1), 010101.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Ajzen, I., & Fishbein, M. (2000). Attitudes and the attitude-behavior relation: Reasoned and automatic processes. *European review of social psychology*, *11*(1), 1-33.

Allen, J. M., & Nimon, K. (2007). Retrospective pretest: a practical technique for professional development evaluation. *Journal of Industrial Teacher Education*, *44*(3), 27-42.

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British journal of social psychology*, *40*(4), 471-499.

Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education*, *90*(5), 536-545.

Bell, C., Gitomer, D., Savage, C., & McKenna, A. H. (2019). A synthesis of research on and measurement of STEM teacher preparation. *American Association for the Advancement of Science. https://aaas-arise. org/wp-content/uploads/2020/01/Bell-Gitomer-Savage-McKenna-A-Synthesis-of-Research-on-and-Measurement-of-STEM-Teacher-Preparation. Pdf*.

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior research methods, 50*, 1953-1959.

Borrego, M., Froyd, J. E., & Hall, T. S. (2010). Diffusion of engineering education innovations: A survey of awareness and adoption rates in US engineering departments. *Journal of Engineering Education, 99*(3), 185-207.

Borda, E., Haskell, T., & Boudreaux, A. (2022). Cross-disciplinary learning: A framework for assessing application of concepts across science disciplines. *Journal of College Science Teaching*, *52*(1), 3-5.

Borrego, M., & Henderson, C. (2014). Increasing the use of evidence-based teaching in STEM higher education: A comparison of eight change strategies. *Journal of Engineering Education, 103*(2), 220-252.

Brewe, E., Sawtelle, V., Kramer, L. H., O'Brien, G. E., Rodriguez, I., & Pamelá, P. (2010). Toward equity through participation in Modeling Instruction in introductory university physics. *Physical Review Special Topics-Physics Education Research*, *6*(1), 010106.

Callens, M. V., Kelter, P., Motschenbacher, J., Nyachwaya, J., Ladbury, J. L., & Semanko, A. M. (2019). Developing and implementing a campus-wide professional development program. *Journal of College Science Teaching*, *49*(2), 68-75.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & cognition*, *42*, 434-447.

Chasteen, S. V., & Chattergoon, R. (2020). Insights from the Physics and Astronomy New Faculty Workshop: How do new physics faculty teach?. *Physical Review Physics Education Research, 16*(2), 020164.

Chasteen, S. V., Chattergoon, R., Prather, E. E., & Hilborn, R. (2016, December). Evaluation methodology and results for the new faculty workshops. AMER ASSOC PHYSICS TEACHERS.

Chiu, M. M., & Kuo, S. W. (2010). From metacognition to social metacognition: Similarities, differences, and learning. *Journal of Education Research*, *3*(4), 321-338.

Cortright, R. N., Collins, H. L., Rodenbaugh, D. W., & DiCarlo, S. E. (2003). Student retention of course content is improved by collaborative-group testing. *Advances in Physiology Education*, *27*(3), 102-108.

Dancy, M., & Henderson, C. (2008, October). Barriers and promises in STEM reform. In *National Academies of Science Promising Practices Workshop* (Vol. 15, pp. 1-17).

DiSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and instruction*, *10*(2-3), 105-225.

Durrant, L. K., Pierson, G., & Allen, E. M. (1985). Group testing and its effectiveness in learning selected nursing concepts. *Journal of the Royal Society of Health*, *105*(3), 107-111.

Drennan, J., & Hyde, A. (2008). Controlling response shift bias: the use of the retrospective pre-test design in the evaluation of a master's programme. *Assessment & Evaluation in Higher Education*, *33*(6), 699-709.

Efu, S. I. (2019). Exams as learning tools: A comparison of traditional and collaborative assessment in higher education. *College teaching*, *67*(1), 73-83.

Elby, A. (2000). What students' learning of representations tells us about constructivism. *The Journal of Mathematical Behavior*, *19*(4), 481-502.

Evans, J. S. B. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, *75*(4), 451-468.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223-241.

Fishbein, M., & Ajzen, I. (2011). *Predicting and changing behavior: The reasoned action approach*. Psychology press.

Fiske, S. T., & Taylor, S. E. (1991). *Social cognition*. Mcgraw-Hill Book Company.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, *19*(4), 25-42.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the national academy of sciences*, *111*(23), 8410-8415.

Gette, C. R., & Kryjevskaia, M. (2019). Establishing a relationship between student cognitive reflection skills and performance on physics questions that elicit strong intuitive responses. *Physical Review Physics Education Research*, *15*(1), 010118.

Gette, C. R., Kryjevskaia, M., Stetzer, M. R., & Heron, P. R. (2018). Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy. *Physical Review Physics Education Research*, *14*(1), 010113.

Gilley, B. H., & Clarkston, B. (2014). Collaborative testing: Evidence of learning in a controlled in-class study of undergraduate students. *Journal of College Science Teaching*, *43*(3), 83-91.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

Goldberg, F. M., Robinson, S., & Otero, V. K. (2008). *Physics & everyday thinking*. It's About Time, Herff Jones Educational Division.

Goos, M., Galbraith, P., & Renshaw, P. (2002). Socially mediated metacognition: Creating collaborative zones of proximal development in small group problem solving. *Educational studies in Mathematics*, *49*, 193-223.

Hazari, Z., Chari, D., Potvin, G., & Brewe, E. (2020). The context dependence of physics identity: Examining the role of performance/competence, recognition, interest, and sense of belonging for lower and upper female physics undergraduates. *Journal of Research in Science Teaching*, *57*(10), 1583-1607.

Heckler, A. F. (2011). 8 The Ubiquitous Patterns of Incorrect Answers to Science Questions: The Role of Automatic, Bottom-up Processes. *Psychology of Learning and Motivation-Advances in Research and Theory, 55,* 227.

Heckler, A. F., & Bogdan, A. M. (2018). Reasoning with alternative explanations in physics: The cognitive accessibility rule. *Physical Review Physics Education Research*, *14*(1), 010120.

Heller, P., Keith, R., & Anderson, S. (1992). Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving. *American journal of physics, 60*(7), 627-636.

Heller, P., & Hollabaugh, M. (1992). Teaching problem solving through cooperative grouping. Part 2: Designing problems and structuring groups. *American journal of Physics*, *60*(7), 637-644.

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature. *Journal of research in science teaching*, *48*(8), 952-984.

Henderson, C., Beach, A. L., & Finkelstein, N. (2012). Four categories of change strategies for transforming undergraduate instruction. *Transitions and transformations in learning and education*, 223-245.

Henderson, C., & Dancy, M. H. (2007). Barriers to the use of research-based instructional strategies: The influence of both individual and situational characteristics. *Physical Review Special Topics-Physics Education Research*, *3*(2), 020102.

Henderson, Charles, and Melissa H. Dancy. "Physics faculty and educational researchers: Divergent expectations as barriers to the diffusion of innovations." *American Journal of Physics* 76.1 (2008): 79-91.

Henderson, C., Dancy, M., & Niewiadomska-Bugaj, M. (2012). Use of research-based instructional strategies in introductory physics: Where do faculty leave the innovation-decision process?. *Physical Review Special Topics-Physics Education Research*, *8*(2), 020104.

Heron, P. R. (2017). Testing alternative explanations for common responses to conceptual questions: An example in the context of center of mass. *Physical Review Physics Education Research*, *13*(1), 010131.

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, memory, and cognition*, *34*(5), 1191.

Thompson, J., Gibson, T., & Higgens, D. (1996). Clustal W version 1.6. *European Molecular Biology Laboratory, Meyerhofstrasse*, *1*.

Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pretest. *American Journal of Evaluation*, *26*(4), 501-517.

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*(2), 144.

Hsu, L., Brewe, E., Foster, T. M., & Harper, K. A. (2004). Resource letter RPS-1: Research in problem solving. *American journal of physics, 72*(9), 1147-1156.

Jang, H., Lasry, N., Miller, K., & Mazur, E. (2017). Collaborative exams: cheating? Or learning?. *American Journal of Physics, 85*(3), 223-227.

Johnson-Laird, P. N. (2006). Mental models, sentential reasoning, and illusory inferences. In *Advances in psychology* (Vol. 138, pp. 27-51). North-Holland.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Khatri, R., Henderson, C., Cole, R., Froyd, J. E., Friedrichsen, D., & Stanford, C. (2015). Characteristics of well-propagated undergraduate STEM teaching innovations. In *Physics Education Research Conference*.

Knierim, K., Turner, H., & Davis, R. K. (2015). Two-stage exams improve student learning in an introductory geology course: Logistics, attendance, and grades. *Journal of Geoscience Education, 63*(2), 157-164.

Kryjevskaia, M., Heron, P. R., & Heckler, A. F. (2021). Intuitive or rational? Students and experts need to be both. *Physics Today, 74*(8), 28-34.

Kryjevskaia, M., & Stetzer, M. R. (2013, January). Examining inconsistencies in student reasoning approaches. In *AIP Conference Proceedings* (Vol. 1513, No. 1, pp. 226-229). American Institute of Physics.

Kryjevskaia, M., Stetzer, M. R., & Grosz, N. (2014). 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), 020109.

Kryjevskaia, M., Stetzer, M. R., & Le, T. K. (2015). Failure to engage: Examining the impact of metacognitive interventions on persistent intuitive reasoning approaches. In *Proceedings of 2014 physics education research conference. AAPT, College Park* (pp. 143-146).

Kryjevskaia, M., Stetzer, M. R., Lindsey, B. A., McInerny, A., Heron, P. R., & Boudreaux, A. (2020). Designing research-based instructional materials that leverage dual-process theories of reasoning: Insights from testing one specific, theory-driven intervention. *Physical Review Physics Education Research*, *16*(2), 020140.

Lamb, T. A., & Tschillard, R. (2005). Evaluating learning in professional development workshops: Using the retrospective pretest. *Journal of Research in Professional Learning*, *1*, 1-9.

Lambiotte, J. G., Dansereau, D. F., Rocklin, T. R., Fletcher, B., Hythecker, V. I., Larson, C. O., & O'Donnell, A. M. (1987). Cooperative learning and test taking: Transfer of skills. *Contemporary Educational Psychology*, *12*(1), 52-61.

Laws, P. W. (2004). *The Physics Suite: Workshop Physics Activity Guide, Module 2: Mechanics II*. John Wiley & Sons.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of behavioral decision making*, *25*(4), 361-381.

Lindsey, B. A., Heron, P. R., & Shaffer, P. S. (2012). Student understanding of energy: Difficulties related to systems. *American Journal of Physics*, *80*(2), 154-163.

Lindsey, B. A., Nagel, M. L., & Savani, B. N. (2019). Leveraging understanding of energy from physics to overcome unproductive intuitions in chemistry. *Physical Review Physics Education Research*, *15*(1), 010120.

Lindsey, B. A., Stetzer, M. R., Speirs, J. C., Ferm Jr, W. N., & Van Hulten, A. (2023). Investigating student ability to follow reasoning chains: The role of conceptual understanding. *Physical Review Physics Education Research*, *19*(1), 010128.

Lising, L., & Elby, A. (2005). The impact of epistemology on learning: A case study from introductory physics. *American Journal of Physics*, *73*(4), 372-382.

Lusk, M., & Conklin, L. (2003). Collaborative testing to promote learning. *Journal of Nursing Education*, *42*(3), 121-124.

Lyman, F. (1981). Strategies for Reading Comprehension Think Pair Share. *Unpublished University of Maryland Paper.(Online)(http://www. roe13. k12. il. us/Services/KeriKorn/BDA/ThinkPairShare. pdf) diakses*, *12*.

Lynch, K. B. (2002). When you don't know what you don't know: Evaluating workshops and training sessions using the retrospective pretest methods. *Arlington, VA*.

Mamede, S., van Gog, T., Moura, A. S., de Faria, R. M., Peixoto, J. M., Rikers, R. M., & Schmidt, H. G. (2012). Reflection as a strategy to foster medical students' acquisition of diagnostic competence. *Medical education*, *46*(5), 464-472.

E. Mazur, *Peer Instruction: A User's Manual* (Prentice-Hall, Upper Saddle River, NJ, 1997)

Mazur, E., & Pedigo, D. (2014). *Principles & practice of physics*.

McInerny, A., & Kryjevskaia, M. (2020, September). Investigating a collaborative group exam as an instructional tool to address student reasoning difficulties that remain even after instruction. In *Proceedings of the 2020 Physics Education Research Conference*.

McInerny, A., Kryjevskaia, M., & Leontyev, A. (2021, October). Examining the efficacy of a professional development assessment tool. In *2021 Physics Education Research Conference (PERC)* (pp. 283-288).

McDermott, L. C., & Redish, E. F. (1999). Resource letter: PER-1: Physics education research. *American journal of physics*, *67*(9), 755-767.

McDermott, L. C., & Shaffer, P. S. (2002). *Tutorials in introductory physics* (pp. 1-245). Upper Saddle River, NJ: Prentice Hall.

McPadden, D., Brewe, E., Monsalve, C., & Sawtelle, V. (2020). Productive faculty resources activated by curricular materials: An example of epistemological beliefs in University Modeling Instruction. *Physical Review Physics Education Research*, *16*(2), 020158.

Menekse, M., Stump, G. S., Krause, S., & Chi, M. T. (2013). Differentiated overt learning activities for effective instruction in engineering classrooms. *Journal of Engineering Education*, *102*(3), 346-374.

Meltzer, D. E., & Thornton, R. K. (2012). Resource letter ALIP–1: active-learning instruction in physics. *American journal of physics*, *80*(6), 478-496.

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision making*, *13*(3), 246-259.

Mikula, B. D., & Heckler, A. F. (2017). Framework and implementation for improving physics essential skills via computer-based practice: Vector math. *Physical Review Physics Education Research*, *13*(1), 010122.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in cognitive sciences*, *14*(10), 435-440.

National Research Council, Division of Behavioral, Center for Education, & Committee on the Study of Teacher Preparation Programs in the United States. (2010). *Preparing teachers: Building evidence for sound policy*. National Academies Press.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175-220.

Novak, G. M., Patterson, E. T., Gavrin, A., & Enger, R. C. (1998, May). Just-in-Time Teaching: Active learner pedagogy with WWW. In *IASTED International Conference on Computers and Advanced Technology in Education* (Vol. 1998, pp. 27-30).

Olson, S., & Riordan, D. G. (2012). Engage to excel: producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. Report to the president. *Executive office of the president*.

Osman, M., & Stavy, R. (2006). Development of intuitive rules: Evaluating the application of the dual-system framework to understanding children’s intuitive reasoning. *Psychonomic Bulletin & Review, 13*, 935-953.

Pan, S. C., Cooke, J., Little, J. L., McDaniel, M. A., Foster, E. R., Connor, L. T., & Rickard, T. C. (2019). Online and clicker quizzing on jargon terms enhances definition-focused but not conceptually focused biology exam performance. *CBE—Life Sciences Education*, *18*(4), ar54.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition?. *Behavior research methods, 48*, 341-348.

Perkins, D. (1995). Outsmarting IQ: The emerging science of learnable intelligence, New York, NY: Free Press.

Perkins, K., Adams, W., Dubson, M., Finkelstein, N., Reid, S., Wieman, C., & LeMaster, R. (2006). PhET: Interactive simulations for teaching and learning physics. *The physics teacher, 44*(1), 18-23.

Prather, E. E., Slater, T. F., Adams Jeffrey, P., Brissenden, G., Dostal, J. A., & Wallace, C. S. (2013). Lecture-tutorials for introductory astronomy. *(No Title)*.

Raidl, M., Johnson, S., Gardiner, K., & Denham, M. (2004). Use retrospective surveys to obtain complete data sets and measure impact in extension programs. *The Journal of Extension*, *42*(2), 13.

Ramlo, S. (2008). Validity and reliability of the force and motion conceptual evaluation. *American Journal of Physics*, *76*(9), 882-886.

Reichenbach, R. S. D. (2023). *Measuring the Middle: A Longitudinal Examination of Instructional Change During and Following a Two-Year Professional Development Program for University Faculty* (Doctoral dissertation, North Dakota State University).

Rieger, G. W., & Heiner, C. E. (2014). Examinations that support collaborative learning: The students' perspective. *Journal of College Science Teaching*, *43*(4), 41-47.

Sabella, M. S., & Cochran, G. L. (2004, September). Evidence of intuitive and formal knowledge in student responses: examples from the context of dynamics. In *AIP Conference Proceedings* (Vol. 720, No. 1, pp. 89-92). American Institute of Physics.

Sawilowsky, S. S. (2009). New effect size rules of thumb. Journal of modern applied statistical methods, 8, 597-599.

Sawtelle, V., Brewe, E., & Kramer, L. H. (2012). Exploring the relationship between self-efficacy and retention in introductory physics. *Journal of research in science teaching*, *49*(9), 1096-1121.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological review*, *112*(3), 610.

Schwarz, N., Bless, H., Wänke, M., & Winkielman, P. (2004). Accessibility revisited. In *Foundations of social cognition* (pp. 51-77). Psychology Press.

Semanko, A. M., & Ladbury, J. L. (2020). Using the reasoned action approach to predict active teaching behaviors in college STEM courses. *Journal for STEM Education Research, 3*(3), 387-402.

Shaffer, P. S., & McDermott, L. C. (1992). Research as a guide for curriculum development: An example from introductory electricity. Part II: Design of instructional strategies. *American journal of physics*, *60*(11), 1003-1013.

Sharma, M. D., Johnston, I. D., Johnston, H., Varvell, K., Robertson, G., Hopkins, A., ... & Thornton, R. (2010). Use of interactive lecture demonstrations: A ten year study. *Physical Review Special Topics-Physics Education Research*, *6*(2), 020119.

Shirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection. *Cognitive science*, *26*(4), 469-501.

Siegel, M. A. (2012). Filling in the distance between us: Group metacognition during problem solving in a secondary education course. *Journal of Science Education and Technology, 21*, 325-341.

Simon, H. A. (1992). What is an “explanation” of behavior?. *Psychological science*, *3*(3), 150-161.

Smith, Michelle K., et al. "Why peer discussion improves student performance on in-class concept questions." *Science* 323.5910 (2009): 122-124.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE—Life Sciences Education, 12*(4), 618-627.

Smith, T. I., & Wittmann, M. C. (2008). Applying a resources framework to analysis of the Force and Motion Conceptual Evaluation. *Physical Review Special Topics-Physics Education Research, 4*(2), 020101.

Sokoloff, D. R., & Thornton, R. K. (1997, March). Using interactive lecture demonstrations to create an active learning environment. In *AIP Conference Proceedings* (Vol. 399, No. 1, pp. 1061-1074). American Institute of Physics.

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423-444.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. Yale University Press.

Speirs, J. C. (2019). *New Methodologies for Examining and Supporting Student Reasoning in Physics*. The University of Maine.

Speirs, J. C., Leuteritz, R., Lê, T. K., Deng, R., & Ell, S. W. (2023). 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.

Speirs, J. C., Stetzer, M. R., Lindsey, B. A., & Kryjevskaia, M. (2021). Exploring and supporting student reasoning in physics by leveraging dual-process theories of reasoning and decision making. *Physical Review Physics Education Research*, *17*(2), 020137.

Sprangers, M., & Hoogstraten, J. (1989). Pretesting effects in retrospective pretest-posttest designs. *Journal of applied psychology*, *74*(2), 265.

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision making*, *13*(3), 260-267.

Stearns, S. A. (1996). Collaborative exams as learning tools. *College Teaching*, *44*(3), 111-112.

Stupple, E., Gale, M., & Richmond, C. (2013). Working memory, cognitive miserliness and logic as predictors of performance on the cognitive reflection test. In *Proceedings of the Annual Meeting of the Cognitive Science society* (Vol. 35, No. 35).

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., ... & Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, *117*(12), 6476-6483.

Thompson, J. (2009). *Performance affects: Applied theatre and the end of effect*. Springer.

Thompson, V. A., Evans, J. S. B., & Campbell, J. I. (2018). Matching bias on the selection task: It's fast and feels good. In *New Paradigm Psychology of Reasoning* (pp. 194-215). Routledge.

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive psychology*, *63*(3), 107-140.

Thornton, R. K., & Sokoloff, D. R. (1998). Assessing student learning of Newton's laws: The force and motion conceptual evaluation and the evaluation of active learning laboratory and lecture curricula. *American Journal of Physics*, *66*(4), 338-352.

Thornton, R. K., Kuhl, D., Cummings, K., & Marx, J. (2009). Comparing the force and motion conceptual evaluation and the force concept inventory. *Physical review special topics-Physics education research*, *5*(1), 010105.

Tipler, P. A., & Mosca, G. (2007). *Physics for scientists and engineers*. Macmillan.

Tishman, S., Jay, E., & Perkins, D. N. (1993). Teaching thinking dispositions: From transmission to enculturation. *Theory into practice*, *32*(3), 147-153.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition*, *39*(7), 1275-1289.

Traxler, A., & Brewe, E. (2015). Equity investigation of attitudinal shifts in introductory physics. *Physical Review Special Topics-Physics Education Research*, *11*(2), 020132.

Trigwell, K., & Prosser, M. (2004). Development and use of the approaches to teaching inventory. *Educational Psychology Review*, *16*, 409-424.

Trigwell, K., Prosser, M., & Ginns, P. (2005). Phenomenographic pedagogy and a revised approaches to teaching inventory. *Higher Education Research & Development*, *24*(4), 349-360.

Trowbridge, D. E., & McDermott, L. C. (1981). Investigation of student understanding of the concept of acceleration in one dimension. *American journal of Physics*, *49*(3), 242-253.

Turpen, C., Dancy, M., & Henderson, C. (2016). Perceived affordances and constraints regarding instructors' use of Peer Instruction: Implications for promoting instructional change. *Physical Review Physics Education Research*, *12*(1), 010116.

Vickrey, T., Rosploch, K., Rahmanian, R., Pilarz, M., & Stains, M. (2015). based implementation of peer instruction: A literature review. *CBE—Life Sciences Education, 14*(1), es3.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.

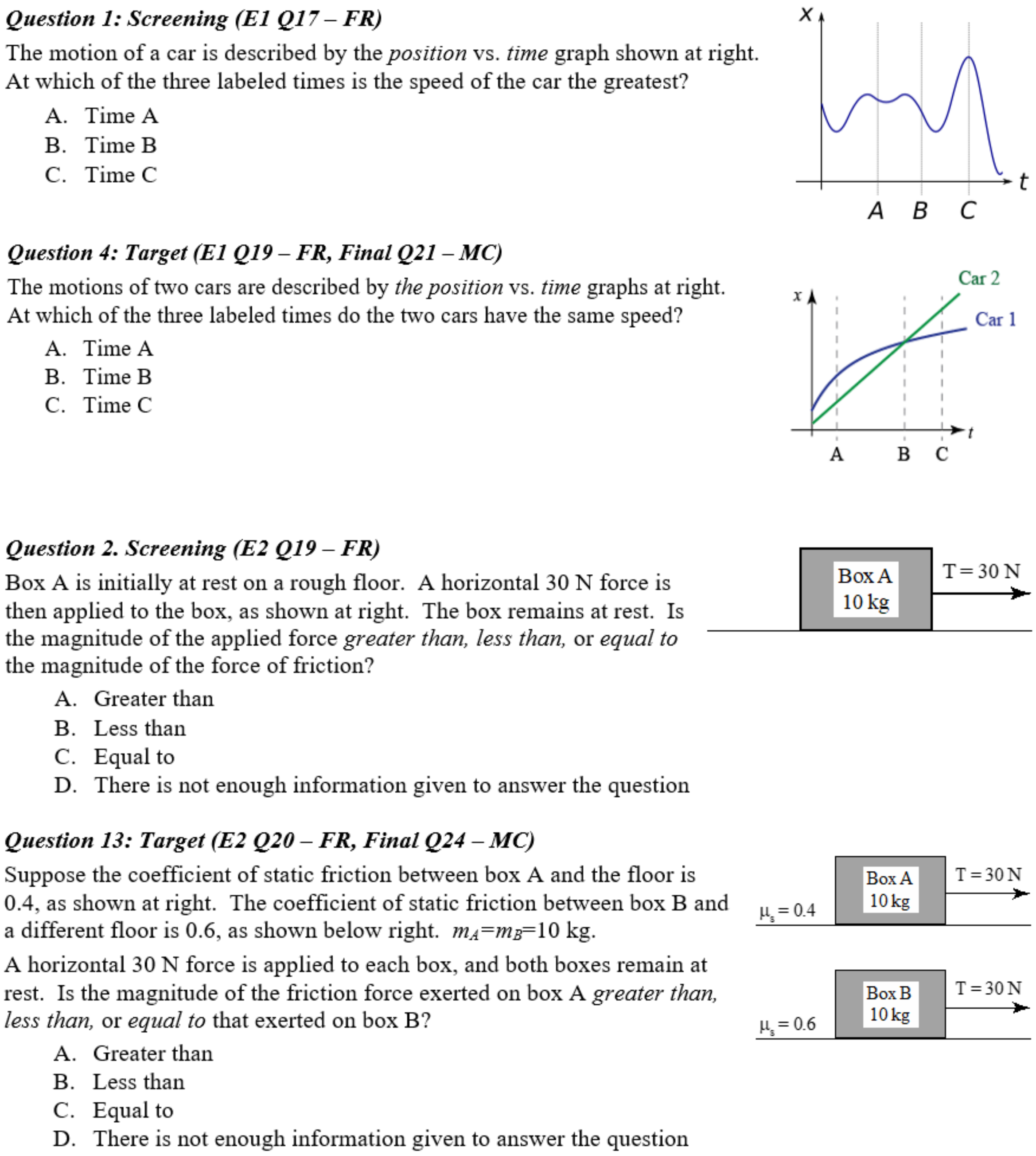
Williams, E. A., Zwolak, J. P., Dou, R., & Brewe, E. (2019). Linking engagement and performance: The social network analysis perspective. *Physical review physics education research*, *15*(2), 020150.

Wood, A. K., Galloway, R. K., & Hardy, J. (2016). Can dual processing theory explain physics students' performance on the Force Concept Inventory?. *Physical Review Physics Education Research*, *12*(2), 023101.

Yuretich, R. F., Khan, S. A., Leckie, R. M., & Clement, J. J. (2001). Active-learning methods to improve student performance and scientific interest in a large introductory oceanography course. *Journal of Geoscience Education*, *49*(2), 111-119.

Zimbardo, P. G., Butler, L. D., & Wolfe, V. A. (2003). Cooperative college examinations: More gain, less pain when students share information and grades. *The Journal of Experimental Education*, *71*(2), 101-125.

APPENDIX: COLLABORATIVE EXAM SCREENING-TARGET QUESTION PAIRS



Question 1: Screening (E1 Q17 – FR)

The motion of a car is described by the *position* vs. *time* graph shown at right. At which of the three labeled times is the speed of the car the greatest?

A. Time A
B. Time B
C. Time C

Question 4: Target (E1 Q19 – FR, Final Q21 – MC)

The motions of two cars are described by *the position* vs. *time* graphs at right. At which of the three labeled times do the two cars have the same speed?

A. Time A
B. Time B
C. Time C

Question 2. Screening (E2 Q19 – FR)

Box A is initially at rest on a rough floor. A horizontal 30 N force is then applied to the box, as shown at right. The box remains at rest. Is the magnitude of the applied force *greater than, less than,* or *equal to* the magnitude of the force of friction?

A. Greater than
B. Less than
C. Equal to
D. There is not enough information given to answer the question

Question 13: Target (E2 Q20 – FR, Final Q24 – MC)

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

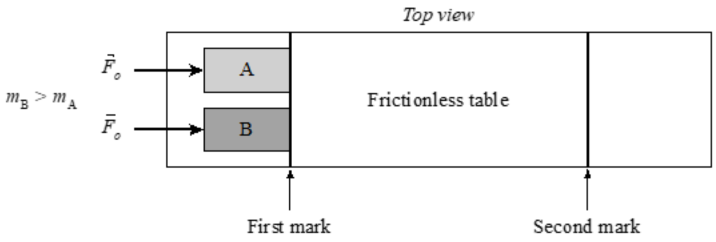
A horizontal 30 N force is applied to each box, and both boxes remain at rest. Is the magnitude of the friction force exerted on box A *greater than, less than,* or *equal to* that exerted on box B?

A. Greater than
B. Less than
C. Equal to
D. There is not enough information given to answer the question

Figure A1. Target screening question pairs.

Question 6: Screening (E3 Q13 – FR)

Two carts, A and B, are initially at rest on a horizontal frictionless table as shown in the top-view diagram at right. A constant force of magnitude F_o is exerted on each cart as it travels between the two marks on the table. Cart B has a greater mass than cart A.



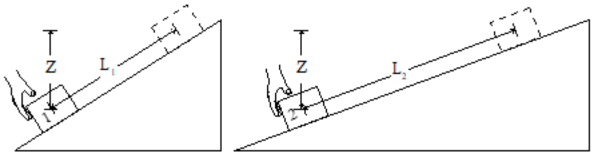
Is the absolute value of the net work done on cart A *($W_{\text{net},A}$) greater than, less than,* or *equal to* the absolute value of the net work done on cart B *($W_{\text{net},B}$)?*

A. $W_{\text{net},A} > W_{\text{net},B}$
B. $W_{\text{net},A} < W_{\text{net},B}$
C. $W_{\text{net},A} = W_{\text{net},B}$
D. Not enough information is given to answer the question

Question 19: Target (E3 Q14 – FR)

Final Q19 - MC

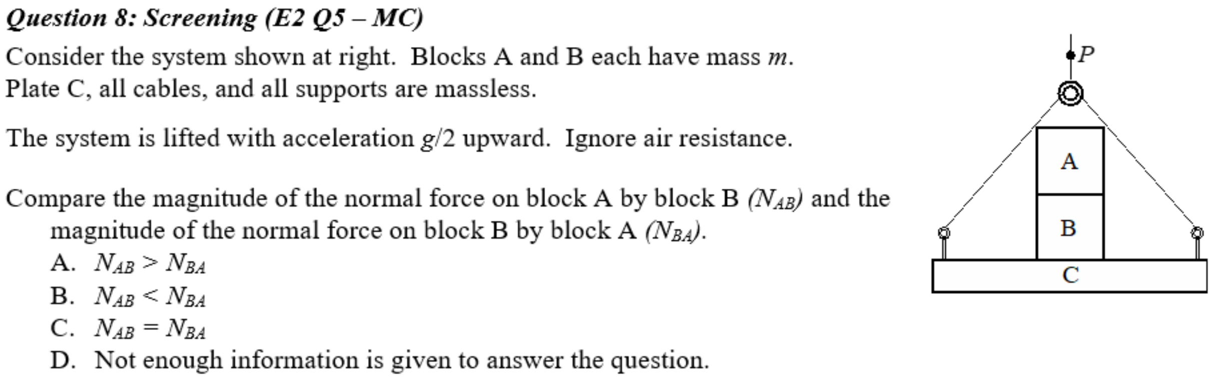
Two identical blocks are pushed up inclines of negligible friction. The hand pushes parallel to the incline with the *same magnitude of force* in each case. Consider the portion of the motion of block 1 during which it moves a distance L_1, and the portion of the motion of block 2 during which it moves a distance L_2. Each block is displaced through the same vertical distance Z.



Is the absolute value of the work by the hand on block 1 *greater than, less than,* or *equal to* the absolute value of the work by the hand on block 2?

A. Greater than
B. Less than
C. Equal to
D. There is not enough information given to answer the question.

Figure A1. Target screening question pairs (continued)



Question 8: Screening (E2 Q5 – MC)

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

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

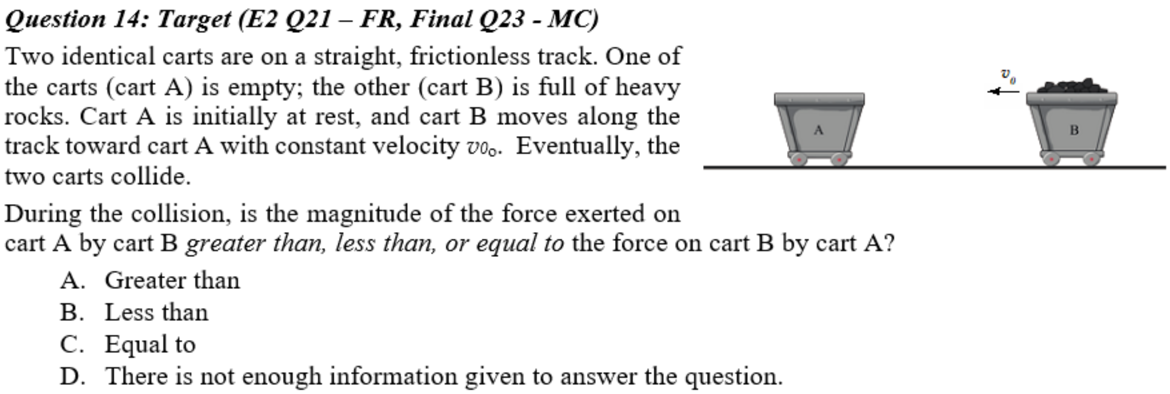
Compare the magnitude of the normal force on block A by block B *(N_{AB})* and the magnitude of the normal force on block B by block A *(N_{BA})*.

A. $N_{AB} > N_{BA}$

B. $N_{AB} < N_{BA}$

C. $N_{AB} = N_{BA}$

D. Not enough information is given to answer the question.



Question 14: Target (E2 Q21 – FR, Final Q23 - MC)

Two identical carts are on a straight, frictionless track. One of the carts (cart A) is empty; the other (cart B) is full of heavy rocks. Cart A is initially at rest, and cart B moves along the track toward cart A with constant velocity $v0_o$. Eventually, the two carts collide.

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

A. Greater than

B. Less than

C. Equal to

D. There is not enough information given to answer the question.

Figure A1. Target screening question pairs (continued)